STIC-ILL

From: Turner, Sharon
Sent: Thursday, June 17, 1999 4:35 PM
To: STIC-ILL
Subject: 09092297

Please RUSH!

Nature 1998, 392:353-58

Virology, 1992, 190:587-596

Mol Cell Biol, 1993, 13:1708-18

Sharon L. Turner, Ph.D.
CM1-8D08 GAU 1645
(703) 308-0056

09,092,297

Th c mplet g n m of th hyperthermophilic bacterium *Aquifex aeolicus*

Gerard Deckert*†, Patrick V. Warren*†, Terry Gaasterland‡, William G. Young*, Anna L. Lenox*, David E. Graham§, Ross Overbeek‡, Marjory A. Snead*, Martin Keller*, Monette Aujay*, Robert Huber‖, Robert A. Feldman*, Jay M. Short*, Gary J. Olsen§ & Ronald V. Swanson*

* *Diversa Corporation, 10665 Sorrento Valley Road, San Diego, California 92121, USA*
‡ *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*
§ *Department of Microbiology, University of Illinois, Urbana, Illinois 61801, USA*
‖ *Lehrstuhl für Mikrobiologie, Universität Regensburg W-8400, Regensburg W-8400, Germany*

***Aquifex aeolicus* was one of the earliest diverging, and is one of the most thermophilic, bacteria known. It can grow on hydrogen, oxygen, carbon dioxide, and mineral salts. The complex metabolic machinery needed for *A. aeolicus* to function as a chemolithoautotroph (an organism which uses an inorganic carbon source for biosynthesis and an inorganic chemical energy source) is encoded within a genome that is only one-third the size of the *E. coli* genome. Metabolic flexibility seems t be reduced as a result of the limited genome size. The use of oxygen (albeit at very low concentrations) as an electron acceptor is allowed by the presence of a complex respiratory apparatus. Although this organism grows at 95 °C, the extreme thermal limit of the Bacteria, only a few specific indications of thermophily are apparent from the genome. Here we describe the complete genome sequence of 1,551,335 base pairs of this evolutionarily and physiologically interesting organism.**

Complete genome sequences have been determined for a number of organisms, including Archaea[1], Bacteria[2–7], and Eukarya[8]. Here we present and explore the genome sequence of *Aquifex aeolicus*. With growth-temperature maxima near 95 °C, *Aquifex pyrophilus* and *A. aeolicus* are the most thermophilic bacteria known. Although isolated and described only recently[9], these species are related to filamentous bacteria first observed at the turn of the century, growing at 89 °C in the outflow of hot springs in Yellowstone National Park[10,11]. The observation of these macroscopic assemblages would later be instrumental in the drive to culture hyperthermophilic organisms[12].

The *Aquificaceae* represent the most deeply branching family within the bacterial domain on the basis of phylogenetic analysis of 16S ribosomal RNA sequences[13,14], although analyses of individual protein sequences vary in their placement of *Aquifex* relative to other groups[15–18]. The genera in this group, *Aquifex* and *Hydrogenobacter*, are thermophilic, hydrogen-oxidizing, microaerophilic, obligate chemolithoautotrophs[9,19–21]. *A. aeolicus* (isolated by R.H. and K. O. Stetter) was cultured at 85 °C under an $H_2/CO_2/O_2$ (79.5:19.5:1.0) atmosphere in a medium containing only inorganic components. *A. aeolicus* does not grow on a number of organic substrates, including sugars, amino acids, yeast extract or meat extract. Unlike its close relative *A. pyrophilus*, *A. aeolicus* has not been shown to grow anaerobically with nitrate as an electron acceptor in the laboratory.

From study of the physiology of the organism, several predictions can be made. As an autotroph, *A. aeolicus* must have genes encoding proteins for one or more modes of carbon fixation and a complete set of biosynthetic genes. As autotrophy is a feature that is distributed throughout the Archaea and Bacteria, most of the associated genes are expected to be of ancient origin and clearly related to those characterized elsewhere. The obligate autotrophy suggests a biosynthetic rather than a degradative character. Oxygen respiration implies the presence of corresponding utilization and tolerance genes. The early divergence of the *Aquificaceae* inferred from ribosomal RNA sequences leads to several questions. Are the machineries for oxygen usage and tolerance homologous to those found in mitochondria and well studied organisms such as *Escherichia coli*, or were they invented separately? If there was far less oxygen when the lineage originated, is there evidence for use of alternative oxidants?

Genome

General features of the *A. aeolicus* genome are listed in Box 1. We classified 1,512 open-reading frames (ORFs) into one of three categories, namely, identified (Table 1), hypothetical, or unknown. Identified ORFs were further classified into one of 57 cellular role categories adapted from Riley[22] (Table 1). The relatively high G + C content of the two 16S-23S-5S rRNA operons (65%) is characteristic of thermophilic bacterial rRNAs[23]. The genome is densely packed: most genes are apparently expressed in polycistronic operons and many convergently transcribed genes overlap slightly. Nonetheless, many genes that are functionally grouped within operons in other organisms, such as the tryptophan or histidine biosynthesis pathways, are found dispersed throughout the *A. aeolicus* genome or appear in novel operons. Even when they encode subunits of the same enzyme, the genes are often separated on the chromosome (for example, *glt*B and *glt*D, the genes encoding the large and small subunits of glutamate synthase). Operon organization of genes for the biosynthesis of amino acids is found in both Archaea and Bacteria but it is not universal in either group. *A. aeolicus* is extreme in that no two amino acid biosynthetic genes are found in the same operon. In contrast, genes required for electron transport, hydrogenase subunits, transport systems, ribosomal subunits, and flagella are often in functionally related operons in *A. aeolicus* (Fig. 1). No introns or inteins (protein splicing elements) were detected in the genome.

A single extrachromosomal element (ECE) was identified during sequencing. Sequence redundancy for the total project was calculated to be 4.83. The ECE, however, is significantly over-represented

† Present addresses: Codex Bioinformatics Services, PO Box 90273, San Diego, California 92169, USA (G.D.); Department of Bioinformatics, SmithKline Beecham Pharmaceuticals, Collegeville, Philadelphia 19426, USA (P.V.W.)

relative to the chromosome; when calculated independently for the final assemblies, redundancies are 4.73 and 8.76 for the chromosome and for the ECE, respectively. The ECE therefore appears to be present at roughly twice the copy number of the chromosome. Although no ORFs on the ECE can be assigned a function with confidence, except for a transposase, two of the predicted proteins show similarity to hypothetical proteins in the *Methanococcus jannaschii* genome[1]. One ORF on the ECE is also present in two identical copies on the *A. aeolicus* chromosome, providing evidence of genetic exchange between the chromosome and the ECE.

Reductive tricarboxylic acid cycle

As an autotroph, *A. aeolicus* obtains all necessary carbon by fixing CO_2 from the environment. An assay for activity of the reductive tricarboxylic acid (TCA) cycle in *A. pyrophilus* cell extracts showed *in vitro* activities for each proposed reaction[24]. The reductive (reverse) TCA cycle fixes two molecules of CO_2 to form acetyl-coenzyme A (acetyl-CoA) and other biosynthetic intermediates[25]. The *A. aeolicus* genome contains genes encoding malate dehydrogenase, fumarate hydratase, fumarate reductase, succinate-CoA ligase, ferredoxin oxidoreductase, isocitrate dehydrogenase, aconitase and citrate synthase, which together could constitute the TCA pathway. There is no biochemical evidence for alternative carbon-fixation pathways in *A. pyrophilus*[24,25] nor is there sequence evidence for such pathways in *A. aeolicus*.

The TCA cycle is vital as it provides the substrates of many biosynthetic pathways. (It is beyond the scope of this report to detail these biosynthetic pathways, but they seem to be typically bacterial, and candidate genes for all or most of the enzymes have been identified in *A. aeolicus*.) The central role of the TCA cycle is emphasized by duplication of many of its constituent genes in *A. aeolicus*. Two genes encode proteins that are similar to malate dehydrogenase (in addition to a lactate dehydrogenase). The fumarate hydratase is split into amino- and carboxy-terminal subunits, as is the case in *M. jannaschii*[1]. Unlinked genes encoding two iron–sulphur proteins of fumarate reductase (alternatively succinate dehydrogenase) accompany a single flavoprotein subunit. Two sets of genes resembling succinate-CoA ligase (both the α- and β-subunits) are present. *A. aeolicus* has two putative operons encoding four-subunit (α, β, γ, δ) 2-acid ferredoxin oxidoreductases; members of this family catalyze reversible carboxylation/decarboxylation of pyruvate, 2-isoketovalerate, or 2-oxoglutarate with varying specificity[26]. These duplicated genes may encode paralogous proteins with unique substrate specificity, as opposed to redundant functions. For example, a paralogue of succinate-CoA ligase may activate citrate with coenzyme A to form citryl-CoA, which citrate synthase can cleave to produce oxaloacetate and acetyl-CoA.

Gluconeogenesis through the Embden–Meyerhof–Parnas pathway

Growing autotrophically, *A. aeolicus* must synthesize pentose and hexose monosaccharides from products of the reductive TCA cycle. Pyruvate produced by pyruvate ferredoxin oxidoreductase or by pyruvate carboxylase (oxaloacetate decarboxylase)[24] may enter the Embden–Meyerhof–Parnas pathway of glycolysis and gluconeogenesis. Genes encoding fructose-1,6-bisphosphatase, an essential gluconeogenic enzyme in *E. coli*, have not been identified in the genomes of the autotrophs *A. aeolicus* or *M. jannaschii*[1], suggesting that an unidentified pathway may exist. The *A. aeolicus* genome also encodes enzymes of the pentose-phosphate pathway and enzymes for glycogen synthesis and catabolism. We found neither (phospho) gluconate dehydrase nor 2-keto-3-deoxy-(6-phospho)gluconate aldolase of the Entner–Doudoroff pathway.

Respiration

Aquifex species are able to grow by using oxygen concentrations as low as 7.5 p.p.m. (R.H. and K. O. Stetter, unpublished observations). The enzymes for oxygen respiration are similar to those of other bacteria: ubiquinol cytochrome *c* oxidoreductase (bc_1 complex), cytochrome *c* (three different genes) and cytochrome *c* oxidase (with two different subunit I genes and two different subunit II genes). The alternative system, with cytochrome *bd* ubiquinol oxidase, is also present. Clearly, the *Aquifex* lineage did not independently invent oxygen respiration. This leaves at least three possibilities: consistent with the ability of *Aquifex* to use very low levels of oxygen, the oxygen-respiration system was highly developed when oxygen had only a small fraction of its present concentration before the advent of oxygenic photosynthesis; contrary to what is implied by the 16S phylogeny, the lineage including *Aquifex* originated after the rise in atmospheric oxygen; or oxygen respiration developed once, and was then laterally transferred among bacterial lineages and acquired by *Aquifex*.

Many other oxidoreductases are present in addition to those obviously involved in oxygen respiration. The physiological role of most of these oxidoreductases is unknown or ambiguous, but two deserve comment. There is a putative nitrate reductase in the genome, although *A. aeolicus* has not been observed to perform NO_3^- respiration, unlike the closely related *A. pyrophilus*. The nitrate reductase gene is adjacent to a nitrate transporter, and may be involved in nitrogen assimilation rather than respiration. It is also possible that *A. aeolicus* has a latent ability to respire with nitrate but that the conditions required have not been found. Two gene sequences show strong similarities to Rieske proteins, even though the rest of the ubiquinol cytochrome *c* oxidoreductase subunits appear only once in the genome. One of these Rieske protein genes is adjacent to a sulphide dehydrogenase subunit, suggesting a role in sulphur respiration.

Oxidative stress

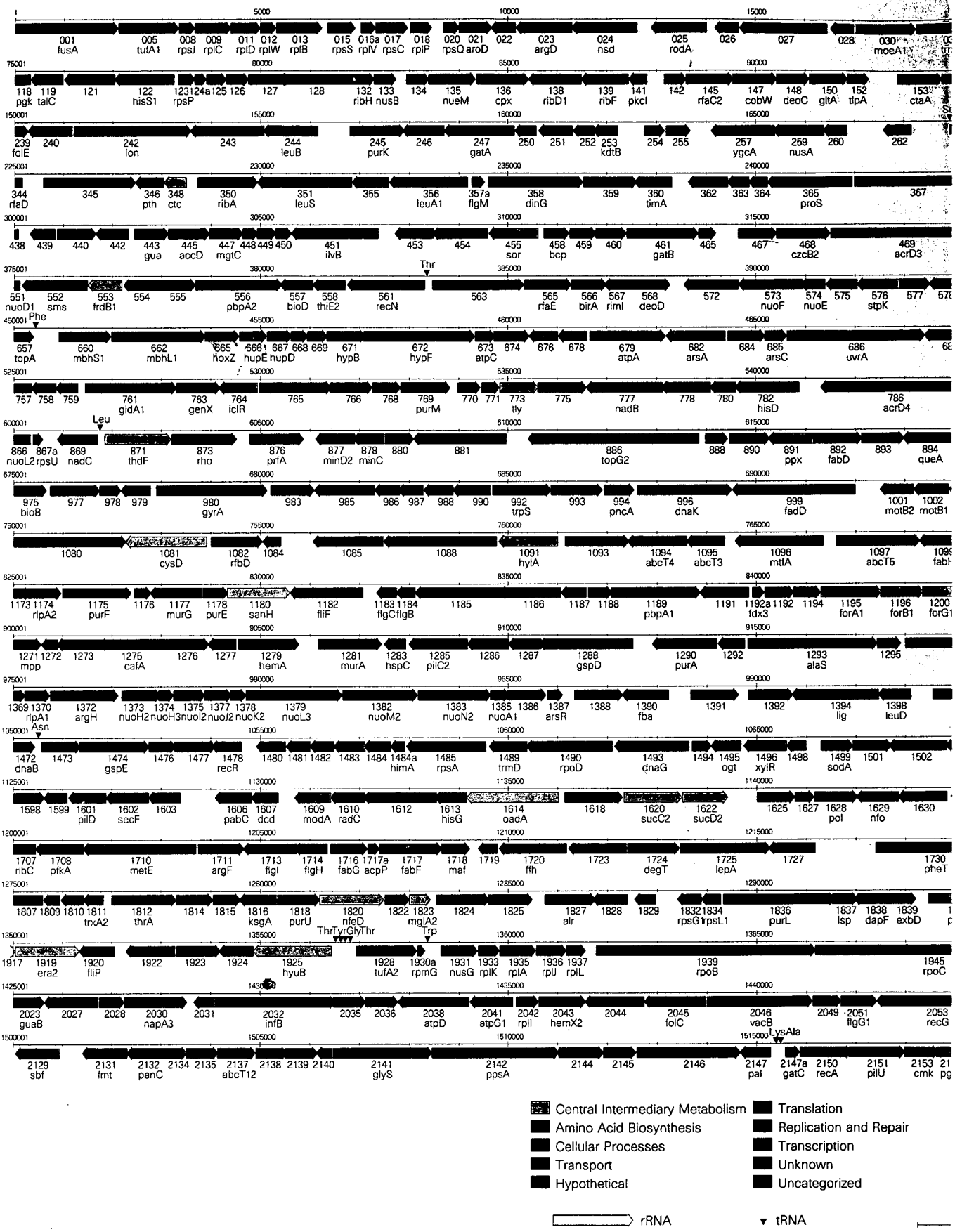
A. aeolicus grows optimally under microaerophilic conditions and consequently possesses various protective enzymes to counter reactive oxygen species, particularly superoxide and peroxide. The genome contains three genes encoding superoxide dismutases, two of the copper/zinc family and one of the iron/manganese family. The latter has also been noted in *A. pyrophilus*[27]. One of the copper/zinc superoxide dismutase genes is located in a large gene cluster encoding formate dehydrogenase.

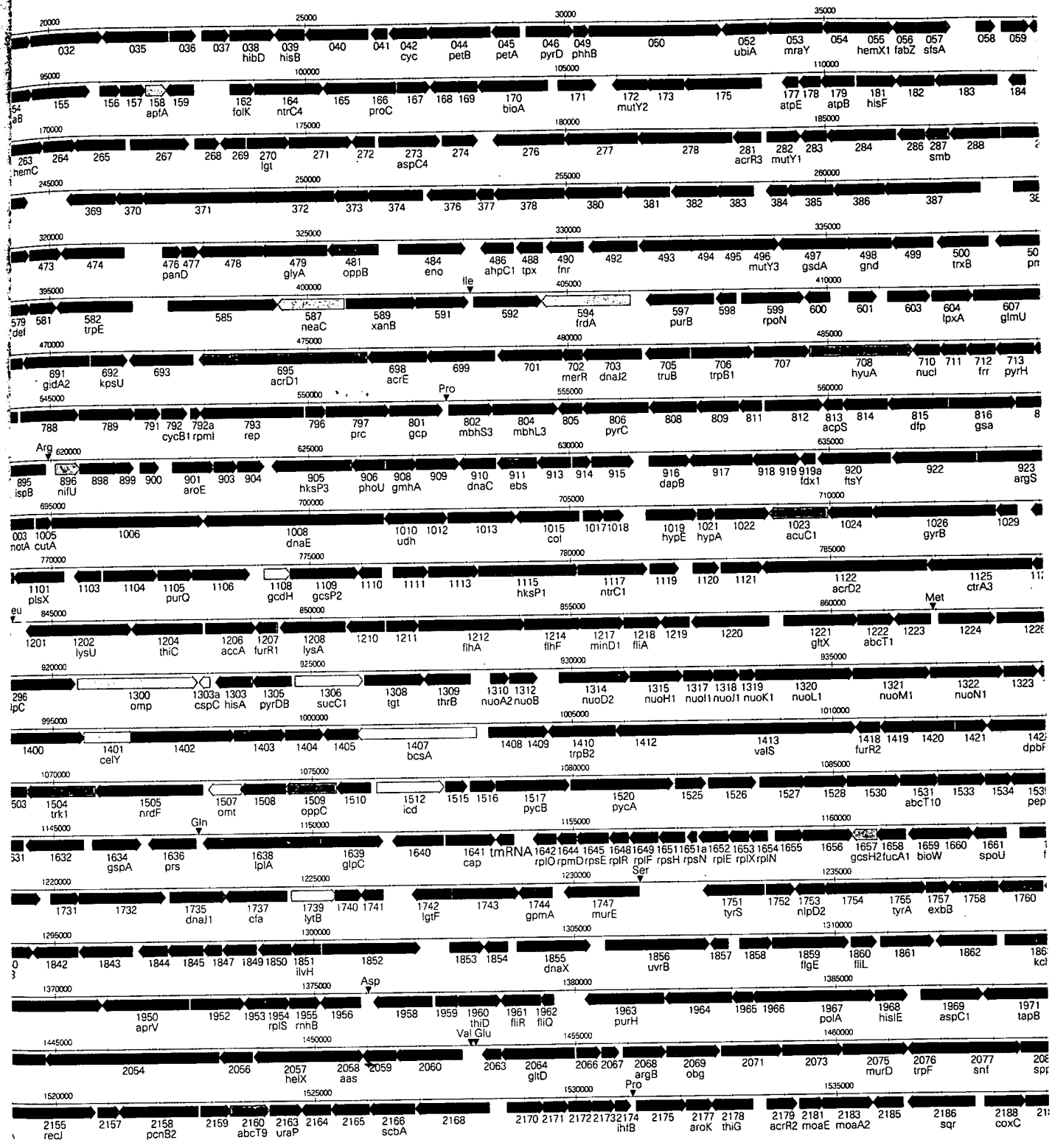
No catalase genes were identified. There are several genes in the genome that might encode proteins that catalyze the detoxification of H_2O_2, including cytochrome *c* peroxidase, thiol peroxidase, and two alkyl hydroperoxide reductase genes. All of these enzymes require an exogenous reductant and therefore do not evolve O_2. However, treatment of *A. pyrophilus*[9] or *A. aeolicus* biomass with H_2O_2 results in the rapid evolution of gas bubbles. This catalase activity may result from a novel enzyme that cannot yet be identified by sequence similarity.

Motility

Like *A. pyrophilus*[9], *A. aeolicus* is motile and possesses monopolar polytrichous flagella. More than 25 genes encoding proteins involved in flagellar structure and biosynthesis have been identified in *A. aeolicus* (Box 1). However, no homologues of the bacterial chemotaxis system were identified. In enteric bacteria, membrane-bound receptors bind chemoattractants and repellents and mod-

Figure 1 Linear map of the *A. aeolicus* circular chromosome. Genes are shown as arrows which denote the direction of transcription and are coloured to denote functional categorization according to the key below the figure. The sequences of the two rRNA gene clusters are identical. Here, the first base of the coding sequence of *fusA* was arbitrarily assigned as base number 1 as no origin of replication has been identified. ORF numbers are discontinuous because some ORFs representing 100 amino acids or more are not predicted to be coding and are not shown.


Central Intermediary Metabolism
Amino Acid Biosynthesis
Cellular Processes
Transport
Hypothetical
Translation
Replication and Repair
Transcription
Unknown
Uncategorized
rRNA
tRNA


Cofactor Biosynthesis
Proteases
Energy Metabolism
Regulation
Purines, Pyrimidines, Nucleotides and Nucleosides
Cell Envelope
Lipid metabolism
Extrachromosomal Element:
5 kb


40000
45000
50000
55000
061 nog
062 speE
063 rpsF
064a ssb
064b rpsR
064c
065
066
067 dmsB
069 rplQ
070 rpoA
072 rpsD
073 rpsK
074 rpsM
075a infA
076 map
078 kad
079 secY
081 aroC
082
083
085 kdsA
086 modC
087
088
090
091
093 hth
115000
120000
125000
130000
185 tagD1
186 aldH1
192 hslU
194
196 trpD1
197
199
200
202
203 atpG2
204
206 nirB
207 cobA
209 trpD2
211 fhp
212 cynS
213 glnBi
215 nasA
190000
195000
200000
205000
91
293
294
296
297 abcT8
298 serS
301 glmS
302
303
305 ileS
308 mutS1
311
313
314
316 hksP2
319 phoB
265000
270000
275000
280000
8
389
390
391
392
394
395 hdrA
397
398 hdrC
400 hdrB
401
402
403
404
406
407
408
409 pyrB
410 carA
411 pcnB1
413 abcT7
340000
345000
350000
355000
1 iu
503
504 otnA'
505 otnA
506
507
509
510
511 murB2
512
515 mtfB
516 mtfC
518 spsK
519 rfe
520 murB1
521 ddlA
522
523 ftsA
525 ftsZ
5 m
415000
420000
425000
430000
608 thrC1
609
613 deaD
615
616
618 pfpI
619 yfeA
620 abcT2
621
622
624 mrcA
626 bioF
627
628
629 xcpC
632
490000
495000
500000
505000
715 tsf
716
717 glgP
718 mpg
720
721 glgA
722 glgB
723 malM
724 ctrA1
727 ldhA
728 speC
729
73 ccd
565000
570000
575000
580000
Ala
Ile
18
819 lnt
820
821 murF
5S2
23S2
16S2
831 exsB
832 surE
833 flgA
834 flgG2
835 nox
836
837 ilvD
841 trm1
843
844 spoT
640000
645000
650000
655000
Ser
924 rnpH
926
928
929 napA2
931
932 dnaQ
933
935
936 ftsH
940 leuC
943
944 gcsH
945 glyQ
946 rnc
948 hemH
950
951 pheA
952 rplT
953 pheS
955 lepB
957
715000
720000
725000
730000
Sel
Val
1030 selD
1031 selA
1033 selB
1035
1036
1037
1038 lysR2
1039 fdoG
1046 fdoH
1049 fdoI
1050 soxC1
1051 fdhE
1052 gcsH1
1053 nifS1
1054
1055 pstB
1056
1057
1058 acrR1
790000
795000
800000
805000
27
1129
1130 sufI
1132 czcB1
1133
1134 proB
1137
1138 rpiB
1139 ftsW
1140 argG
1141
1142
1143 dapA
1144 pabB
1145 dhaT
1148
1151
865000
870000
875000
880000
1229 mfiT2
1230
1231 dmsC
1232 dmsB
1234 dmsA
1236
1237 cysG
1238
1239
1240 adh2
1241
1242 mutS2
1244
1245 ilvC
1247
1248
1249 cds
1250
1253
12
940000
945000
950000
955000
Leu
324
1326 mobB
1327
1328
1329 moeB
1330 gltP
1331 czcB2
1332
1333
1334 pyrG
1335 nse
1336
1337 clpX
1339 clpP
1340 tig
1342 gph
1343
1345
1347
1015000
1020000
1025000
1030000
2
1423 pdxJ
1424 hemF
1426
1427 lpxB
1428
1429 metF
1432 ppdD1
1433 fimZ
1434 ppdD2
1435 ppdD3
1436
1439
1441 oppA
1442
1444
1445 ctrA2
1446
1090000
1095000
1100000
1105000
5
Q
1536 aroA
1539 fliN
1540 gcpE
1542
1543 rfaC1
1545
1546 kpsF
1547 ppa
1548 trpA
1549
1550 cycB2
1552 fabI
1554
1556 cysM
1558
1560 mglA1
1562
1563 abcT13
1565 gltB
1165000
1170000
1175000
1180000
Pro
662 lgK
1663 flgL
1664 accC1
1665 mdh2
Lys
Arg
1666
1667 thrS
1668
1669
1670
1671 hslV
1672 clpB
1677 aspS
1679 fumX
1680
1681 amiB
1682
1240000
1245000
1250000
1255000
His
1761
1763
1765 tktA
1767 rpsT
1768
1769 dld1
1770 leuS'
1771
1772 envA
1773
1774 glnE
1776
1777 infC
1780 fumB
1782 mdh1
1315000
1320000
1325000
1330000
3
1866 asd
1869
1870 secA
1871 pmbA
1873
1875
1877
1878 rplM
rpsI
1879 argC
1882 dnaN
1884
1886 sbcD
1887 mesJ
1888 sucD1
1889 nadA
1890 tsnR
1
1390000
1395000
1400000
1405000
1973 panB
1974
1975
1977 masA
1978
1979 fucA2
1980
1982
1983 imp2
1985
1986
1987
1988 tolQ
1989
1990 pgmA
1991 htpX
1993
1994 era1
1996
1997
1998 fliC
2000
2001 fliD
1465000
1470000
1475000
1480000
0
uA
2082
2084 hisC
2085
2086
2087
2088
2090 leuA2
2093 dsbC
2094 aspC2
2095 dksA
2096 floX
2098
2099 pepA
2101 carB
2102 prmA
2103 acs'
2104 acs
2106 ssf
1540000
1545000
1550000
1551335
39
2190 coxB
2191 coxA1
2192 coxA2
2194
2195 bacA
2196
2197
2199 mopB
2200 mopA
2203
2204 ymxG
15000
20000
25000
30000
ece015
ece017
ece018
ece019
ece020
ece021
ece022
ece023
ece024
ece025
ece026
ece027
ece028
ece029
ece031
ece032
e

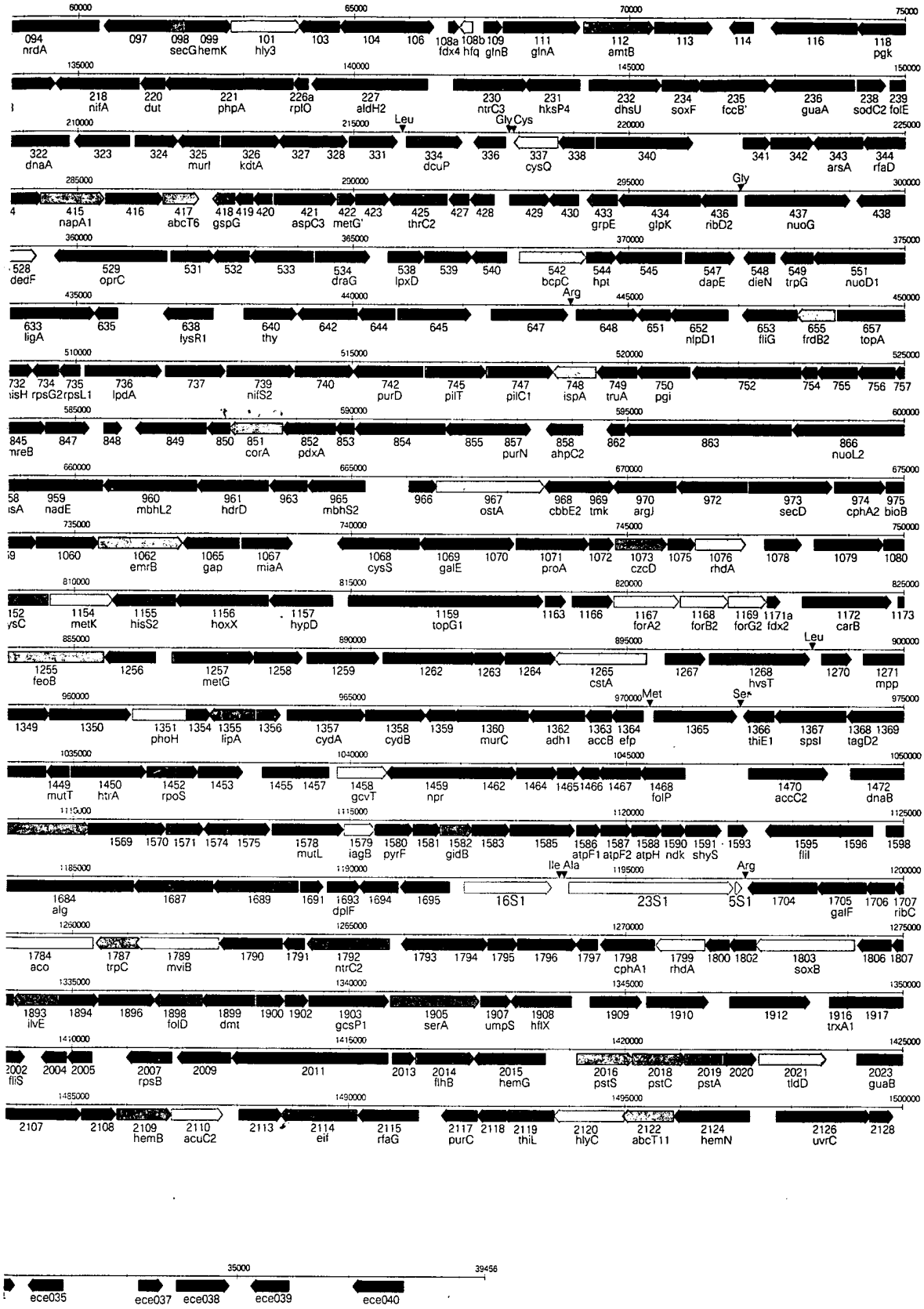

60000 65000 70000 75000
094 nrdA 097 098 secG 099 hemK 101 hly3 103 104 106 108a fdx4 108b hfq 109 glnB 111 glnA 112 amtB 113 114 116 118 pgk
135000 140000 145000 150000
218 nifA 220 dut 221 phpA 226a rplO 227 aldH2 230 ntrC3 231 hksP4 232 dhsU 234 soxF 235 fccB' 236 guaA 238 sodC2 239 folE
210000 215000 220000 225000
Leu Gly Cys
322 dnaA 323 324 325 murI 326 kdtA 327 328 331 334 dcuP 336 337 cysQ 338 340 341 342 343 arsA 344 rfaD
285000 290000 295000 300000
Gly
4 415 napA1 416 417 abcT6 418 gspG 419 420 421 aspC3 422 metG' 423 425 thrC2 427 428 429 430 433 grpE 434 glpK 436 ribD2 437 nuoG 438
360000 365000 370000 375000
528 dedF 529 oprC 531 532 533 534 draG 538 lpxD 539 540 542 bcpC 544 hpt 545 547 dapE 548 dieN 549 trpG 551 nuoD1
435000 440000 445000 450000
Arg
633 ligA 635 638 lysR1 640 thy 642 644 645 647 648 651 652 nlpD1 653 fliG 655 frdB2 657 topA
510000 515000 520000 525000
732 734 735 736 737 739 740 742 745 747 748 749 750 752 754 755 756 757
isH rpsG2 rpsL1 lpdA nifS2 purD pilT pilC1 ispA truA pgi
585000 590000 595000 600000
845 847 848 849 850 851 852 853 854 855 857 858 862 863 866
mreB corA pdxA purN ahpC2 nuoL2
660000 665000 670000 675000
58 959 960 961 963 965 966 967 968 969 970 972 973 974 975
isA nadE mbhL2 hdrD mbhS2 ostA cbbE2 tmk argJ secD cphA2 bioB
735000 740000 745000 750000
9 1060 1062 1065 1067 1068 1069 1070 1071 1072 1073 1075 1076 1078 1079 1080
emrB gap miaA cysS galE proA czcD rhdA
810000 815000 820000 825000
152 1154 1155 1156 1157 1159 1163 1166 1167 1168 1169 1171a 1172 1173
ysC metK hisS2 hoxX hypD topG1 forA2 forB2 forG2 fdx2 carB
885000 890000 895000 900000
Leu
1255 feoB 1256 1257 metG 1258 1259 1262 1263 1264 1265 cstA 1267 1268 hvsT 1270 1271 mpp
960000 965000 970000 975000
Met Ser
1349 1350 1351 phoH 1354 1355 lipA 1356 1357 cydA 1358 cydB 1359 1360 murC 1362 adh1 1363 accB 1364 efp 1365 1366 thiE1 1367 spsI 1368 1369 tagD2
1035000 1040000 1045000 1050000
1449 mutT 1450 htrA 1452 rpoS 1453 1455 1457 1458 gcvT 1459 npr 1462 1464 1465 1466 1467 1468 folP 1470 accC2 1472 dnaB
1110000 1115000 1120000 1125000
1569 1570 1571 1574 1575 1578 mutL 1579 iagB 1580 pyrF 1581 1582 gidB 1583 1585 1586 atpF1 1587 atpF2 1588 atpH 1590 ndk 1591 shyS 1593 1595 flil 1596 1598
1185000 1190000 1195000 1200000
Ile Ala Arg
1684 alg 1687 1689 1691 1693 dplF 1694 1695 16S1 23S1 5S1 1704 1705 galF 1706 1707 ribC
1260000 1265000 1270000 1275000
1784 aco 1787 trpC 1789 mviB 1790 1791 1792 ntrC2 1793 1794 1795 1796 1797 1798 cphA1 1799 rhdA 1800 1802 1803 soxB 1806 1807
1335000 1340000 1345000 1350000
1893 ilvE 1894 1896 1898 folD 1899 dmt 1900 1902 1903 gcsP1 1905 serA 1907 umpS 1908 hflX 1909 1910 1912 1916 trxA1 1917
1410000 1415000 1420000 1425000
2002 fliS 2004 2005 2007 rpsB 2009 2011 2013 2014 flhB 2015 hemG 2016 pstS 2018 pstC 2019 pstA 2020 2021 tldD 2023 guaB
1485000 1490000 1495000 1500000
2107 2108 2109 hemB 2110 acuC2 2113 2114 eif 2115 rfaG 2117 purC 2118 2119 thiL 2120 hlyC 2122 abcT11 2124 hemN 2126 uvrC 2128
35000 39456
ece035 ece037 ece038 ece039 ece040

Table 1 *Aquifex aeolicus* Open Reading Frame Identifications. Gene numbers (Aq) correspond to those in Fig.1. Percentages refer to the identity found in the best FASTA alignment. The percentage of the sequence covered by the alignment is displayed with bullets as follows 20–40% ·, 40–60% · ·, 60–80% · · ·, 80–100% · · · ·

Amino Acid Biosynthesis			
Aromatic amino acids			
Aq1536	aroA	5-enolpyruvylshikimate-3-phosphate synthetase	43.0% ····
Aq081	aroC	chorismate synthase	55.2% ····
Aq021	aroD	3-dehydroquinate dehydratase	33.3% ····
Aq901	aroE	shikimate 5-dehydrogenase	46.1% ····
Aq2177	aroK	shikimate kinase	36.5% ····
Aq951	pheA	chorismate mutase/prephenate dehydratase	44.0% ····
Aq1548	trpA	tryptophan synthase alpha subunit	44.5% ····
Aq706	trpB1	tryptophan synthase beta subunit	68.0% ····
Aq1410	trpB2	tryptophan synthase beta subunit	50.0% ····
Aq1787	trpC	indole-3-glycerol phosphate synthase	43.3% ····
Aq196	trpD1	phosphoribosylanthranilate transferase	45.1% ····
Aq209	trpD2	phosphoribosylanthranilate transferase	24.9% ····
Aq582	trpE	anthranilate synthase component I	50.0% ····
Aq2076	trpF	phosphoribosyl anthranilate isomerase	45.6% ····
Aq549	trpG	anthranilate synthase component II	59.2% ····
Aq1755	tyrA	prephenate dehydrogenase	36.1% ····
Aspartate family			
Aq1866	asd	aspartate-semialdehyde dehydrogenase	54.6% ····
Aq1969	aspC1	aspartate aminotransferase	53.5% ····
Aq2094	aspC2	aminotransferase (AspC family)	55.4% ····
Aq421	aspC3	aminotransferase (AspC family)	43.3% ····
Aq273	aspC4	aminotransferase (AspC family)	48.5% ····
Aq1143	dapA	dihydrodipicolinate synthase	53.1% ····
Aq916	dapB	dihydrodipicolinate reductase	44.2% ····
Aq547	dapE	succinyl-diaminopimelate desuccinylase	25.8% ····
Aq1838	dapF	diaminopimelate epimerase	35.5% ····
Aq1208	lysA	diaminopimelate decarboxylase	47.4% ····
Aq1152	lysC	aspartokinase	52.2% ····
Aq1710	metE	tetrahydropteroyltriglutamate methyltransferase	45.9% ····
Aq1812	thrA	homoserine dehydrogenase	40.4% ····
Aq1309	thrB	homoserine kinase	38.3% ····
Aq608	thrC1	threonine synthase	64.3% ····
Aq425	thrC2	threonine synthase	61.9% ····
Branched-chain family			
Aq451	ilvB	acetolactate synthase large subunit	53.1% ····
Aq1245	ilvC	acetohydroxy acid isomeroreductase	64.3% ····
Aq837	ilvD	dihydroxyacid dehydratase	58.0% ····
Aq1893	ilvE	branched-chain amino acid aminotransferase	40.3% ····
Aq1851	ilvH	acetolactate synthase	53.2% ····
Aq356	leuA1	2-isopropylmalate synthase	52.1% ····
Aq2090	leuA2	2-isopropylmalate synthase	49.9% ····
Aq244	leuB	3-isopropylmalate dehydrogenase	58.7% ····
Aq940	leuC	large subunit of isopropylmalate isomerase	52.3% ····
Aq1398	leuD	3-isopropylmalate dehydratase	56.6% ···
Glutamate family			
Aq2068	argB	acetylglutamate kinase	54.2% ····
Aq1879	argC	N-Acetyl-gamma-glutamylphosphate reductase	40.6% ····
Aq023	argD	N-acetylornithine aminotransferase	49.5% ····
Aq1711	argF	ornithine carbamoyltransferase	46.2% ····
Aq1140	argG	argininosuccinate synthase	54.9% ····
Aq1372	argH	argininosuccinate lyase	46.4% ····
Aq970	argJ	glutamate N-acetyltransferase	39.8% ····
Aq111	glnA	glutamine synthetase	57.6% ····
Aq109	glnB	nitrogen regulatory PII protein	73.2% ····
Aq1774	glnE	glutamate ammonia ligase adenylyl-transferase	28.4% ····
Aq1565	gltB	glutamate synthase large subunit	44.3% ····
Aq2064	gltD	glutamate synthase small subunit gltD	37.7% ····
Aq1071	proA	gamma-glutamyl phosphate reductase	47.9% ····
Aq1134	proB	glutamate 5-kinase	43.2% ····
Aq166	proC	pyrroline carboxylate reductase	35.1% ····
Histidine			
Aq1303	hisA	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase	40.9% ····
Aq039	hisB	imidazoleglycerolphosphate dehydratase	46.4% ····
Aq2084	hisC	histidinol-phosphate aminotransferase	33.7% ····
Aq782	hisD	histidinol dehydrogenase	49.9% ····
Aq181	hisF	HisF (cyclase)	59.9% ····
Aq1613	hisG	ATP phosphoribosyltransferase	40.3% ····
Aq732	hisH	amidotransferase HisH	47.7% ····
Aq1968	hisIE	phosphoribosyl-ATP pyrophosphohydrolase	43.8% ····
Selenocysteine			
Aq1031	selA	L-seryl-tRNA(ser) selenium transferase	42.7% ····
Aq1030	selD	selenophosphate synthase	37.7% ····
Serine family			
Aq1556	cysM	cysteine synthase, O-acetylserine (thiol) lyase B	45.8% ····
Aq479	glyA	serine hydroxymethyl transferase	62.7% ····
Aq1905	serA	D-3-phosphoglycerate dehydrogenase	44.1% ····
Cell Envelope			
Pili and fimbrae			
Aq1433	fimZ	minor pilin	34.9% ··
Aq1432	ppdD1	pilin	40.6% ····
Aq1434	ppdD2	pilin	26.4% ····
Aq1435	ppdD3	pilin	28.2% ···
Lipoproteins and porins			
Aq270	lgt	prolipoprotein diacylglyceryl transferase	30.1% ····
Aq819	lnt	apolipoprotein N-acyltransferase	25.5% ····
Aq652	nlpD1	lipoprotein	25.4% ····
Aq1753	nlpD2	lipoprotein NlpD fragment	43.2% ····
Aq529	oprC	outer membrane protein c	27.2% ····
Aq2147	pal	peptidoglycan associated lipoprotein	35.1% ··
Aq1370	rlpA1	rare lipoprotein A	61.1% ··
Aq1174	rlpA2	rare lipoprotein A	40.6% ···
Aq2166	scbA	adhesion protein	25.7% ····
Aq619	yfeA	adhesion B precursor	28.5% ····
Peptidoglycan			
Aq1827	alr	alanine racemase	33.2% ····
Aq1681	amiB	N-acetylmuramoyl-L-alanine amidase	31.0% ····
Aq2195	bacA	undecaprenol kinase	43.1% ····
Aq1798	cphA1	beta lactamase precursor	25.0% ···
Aq974	cphA2	beta lactamase precursor	29.4% ···
Aq521	ddlA	D-alanine:D-alanine ligase	38.2% ····
Aq301	glmS	glucosamine-fructose-6-phosphate aminotransferase	43.2% ····
Aq607	glmU	UDP-N-acetylglucosamine pyrophosphorylase	37.6% ···
Aq053	mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase	47.5% ····
Aq624	mrcA	penicillin binding protein 1A	33.2% ····
Aq1281	murA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	45.7% ····
Aq520	murB1	UDP-N-acetylenolpyruvoylglucosamine reductase	35.6% ····
Aq511	murB2	UDP-N-acetylenolpyruvoylglucosamine reductase	38.9% ····
Aq1360	murC	UDP-N-acetylmuramate-alanine ligase	46.1% ····
Aq2075	murD	UDP-N-acetylmuramoylalanine-D-glutamate ligase	29.3% ····
Aq1747	murE	UDP-MurNac-tripeptide synthetase	42.9% ····
Aq821	murF	UDP-MURNAC-pentapeptide sythetase	32.3% ····
Aq1177	murG	phospho-N-acetylmuramoyl-pentapeptide-transferase	30.5% ····
Aq325	murI	glutamate racemase	43.4% ····
Aq1189	pbpA1	penicillin binding protein 2	32.2% ····
Aq556	pbpA2	penicillin binding protein 2	30.3% ····
Aq185	tagD1	glycerol-3-phosphate cytidyltransferase	52.0% ····
Aq1368	tagD2	glycerol-3-phosphate cytidyltransferase	67.2% ···
Surface polysaccharides and lipopolysaccharides			
Aq1684	alg	alginate synthesis-related protein	37.2% ··
Aq1641	cap	capsular polysaccharide biosynthesis protein	30.8% ····
Aq1899	dmt	dolichol-phosphate mannosyltransferase	40.2% ····
Aq1772	envA	UDP-3-0-acyl N-acetylglcosamine deacetylase	36.5% ····
Aq1757	exbB	biopolymer transport exbB	48.2% ····
Aq1839	exbD	biopolymer transport ExbD	34.7% ····
Aq1069	galE	UDP-glucose-4-epimerase	54.7% ····
Aq1705	galF	UDP-glucose pyrophosphorylase	47.2% ····
Aq908	gmhA	phosphoheptose isomerase	63.4% ···
Aq085	kdsA	3-deoxy-d-manno-octulosonic acid 8-phosphate synthase	52.0% ····
Aq326	kdtA	3-deoxy-D-manno-2-octulosonic acid transferase	28.9% ····
Aq253	kdtB	lipopolysaccharide core biosynthesis protein	46.5% ····
Aq1546	kpsF	polysialic acid capsule expression protein	45.9% ····
Aq692	kpsU	3-deoxy-manno-octulosonate cytidylyltransferase	41.3% ····
Aq1742	lgtF	beta 1,4 glucosyltransferase	35.2% ····
Aq604	lpxA	acyl-[acyl-carrier-protein]-UDP-N-acetylglucosamine acyltransferase	47.7% ····
Aq1427	lpxB	lipid A disaccharide synthetase	31.6% ····
Aq538	lpxD	UDP-3-O-[3-hydroxymyristoyl] glucosamine N acyltransferase	43.3% ····
Aq718	mpg	mannose-1-phosphate guanyltransferase	34.1% ····
Aq1096	mtfA	mannosyltransferase A	34.3% ···
Aq515	mtfB	mannosyltransferase B	29.0% ····
Aq516	mtfC	mannosyltransferase C	35.9% ····
Aq1335	nse	nucleotide sugar epimerase	45.8% ····
Aq505	otnA	polysaccharide biosynthesis protein	26.9% ····
Aq504	otnA'	polysaccharide biosynthesis protein (fragment)	37.8% ··
Aq1543	rfaC1	ADP-heptose:LPS heptosyltransferase	30.7% ····
Aq145	rfaC2	ADP-heptose:LPS heptosyltransferase	28.1% ····
Aq344	rfaD	ADP-L-glycero-D-manno-heptose-6-epimerase	39.6% ····
Aq565	rfaE	ADP-heptose synthase	44.0% ····
Aq2115	rfaG	glucosyl transferase I	27.1% ····
Aq1082	rfbD	GDP-D-mannose dehydratase	53.2% ····
Aq519	rfe	undecaprenyl-phosphate-alpha-N-acetylglucosaminyltransferase	24.8% ····
Aq1367	spsI	glucose-1-phosphate thymidylyltransferase	30.4% ··
Aq518	spsK	spore coat polysaccharide biosynthesis protein SpsK	49.5% ···
Aq589	xanB	mannose-6-phosphate isomerase/mannose-1-phosphate guanyl transferase	40.9% ····
Cellular Processes			
Cell division			
Aq698	acrE	acriflavin resistance protein AcrE	24.8% ····
Aq1275	cafA	cytoplasmic axial filament protein	28.5% ····
Aq523	ftsA	cell division protein FtsA	31.9% ····
Aq936	ftsH	cell division protein FtsH	51.1% ····
Aq1139	ftsW	cell division protein FtsW	30.8% ···
Aq920	ftsY	cell division protein FtsY	35.2% ···
Aq525	ftsZ	cell division protein FtsZ	48.6% ····
Aq761	gidA1	glucose inhibited division protein A	50.2% ····
Aq691	gidA2	glucose inhibited division protein A	57.5% ····
Aq1582	gidB	glucose inhibited division protein B	39.4% ····
Aq1718	maf	MAF protein	44.9% ····
Aq1887	mesJ	cell cycle protein MesJ	27.7% ····
Aq878	minC	septum site-determining protein MinC	39.4% ··
Aq1217	minD1	septum site-determining protein MinD	33.1% ····
Aq877	minD2	septum site-determining protein MinD	54.5% ····
Aq845	mreB	rod shape determining protein MreB	57.4% ····
Aq025	rodA	rod shape determining protein RodA	37.6% ····
Aq1130	sufI	periplasmic cell division protein (SufI)	28.1% ····
Chaperones			
Aq154	ctaB	cytochrome c oxidase assembly factor	38.8% ····
Aq1735	dnaJ1	chaperone DnaJ	41.3% ····
Aq703	dnaJ2	chaperone DnaJ	45.1% ····
Aq996	dnaK	Hsp70 chaperone DnaK	59.1% ····
Aq433	grpE	heat shock protein GrpE	38.8% ····
Aq192	hslU	chaperone HslU	57.5% ····
Aq1283	hspC	small heat shock protein (class I)	31.0% ····
Aq1991	htpX	heat shock protein X	51.1% ····
Aq2200	mopA	GroEL	64.4% ····
Aq2199	mopB	GroES	56.2% ···
Detoxification			
Aq486	ahpC1	alkyl hydroperoxide reductase	49.2% ····
Aq858	ahpC2	alkyl hydroperoxide reductase	53.4% ····
Aq685	arsC	arsenate reductase	50.0% ····
Aq136	cpx	cytochrome c peroxidase	48.9% ····
Aq1005	cutA	periplasmic divalent cation tolerance protein	47.0% ····
Aq1499	sodA	superoxide dismutase (Fe/Mn family)	34.2% ····
Aq1050	sodC1	superoxide dismutase (Cu/Zn)	39.5% ····
Aq238	sodC2	superoxide dismutase (Cu/Zn)	39.2% ····
Aq488	tpx	thiol peroxidase	39.5% ····
Motility			
Aq833	flgA	flagellar protein FlgA	
Aq1184	flgB	flagellar basal body rod protein FlgB	
Aq1183	flgC	flagellar biosynthesis FlgC	39.4% ····
Aq1859	flgE	flagellar hook protein FlgE	30.8% ····
Aq2051	flgG1	flagellar hook basal-body protein FlgG	32.8% ····
Aq834	flgG2	flagellar hook basal-body protein FlgG	50.4% ····
Aq1714	flgH	flagellar L-ring protein FlgH	31.9% ····
Aq1713	flgI	flagellar P-ring protein FlgI	46.9% ····
Aq1662	flgK	flagellar hook associated protein FlgK	21.9% ····
Aq1663	flgL	flagellar hook associated protein FlgL	27.1% ··
Aq1212	flhA	flagellar export protein	44.0% ····
Aq2014	flhB	flagellar biosynthetic protein FlhB	39.8% ····
Aq1214	flhF	flagellar biosynthesis FlhF	28.7% ····
Aq1998	fliC	flagellin	59.4% ····

Aq2001	fliD	flagellar hook associated protein FliD	24.3% ..
Aq1182	fliF	Flagellar M-ring protein	32.0% ....
Aq653	fliG	flagellar switch protein FliG	35.9% ....
Aq1595	fliI	flagellar export protein	44.6% ....
Aq1860	fliL	flagellar biosynthesis FliL	30.6% ....
Aq1539	fliN	flagellar switch protein FliN	42.9% ...
Aq1920	fliP	flagellar biosynthetic protein FliP	47.7% ....
Aq1962	fliQ	flagellar biosynthesis protein FliQ	45.5% ....
Aq1961	fliR	flagellar biosynthetic protein FliR	29.7% ....
Aq2002	fliS	flagellar protein FliS	30.8% ....
Aq1003	motA	flagellar motor protein MotA	35.0% ....
Aq1002	motB1	flagellar motor protein MotB	36.8% ....
Aq1001	motB2	flagellar motor protein MotB-like	27.5% ....
Secretion			
Aq1720	ffh	signal recognition particle receptor protein	49.1% ....
Aq1288	gspD	general secretion pathway protein D	27.5% ....
Aq1474	gspE	general secretion pathway protein E	48.8% ....
Aq418	gspG	general secretion pathway protein G	50.7% ....
Aq955	lepB	type-I signal peptidase	33.9% ....
Aq1837	lsp	lipoprotein signal peptidase	37.4% ....
Aq1271	mpp	processing protease	28.7% ....
Aq747	pilC1	fimbrial assembly protein PilC	37.4% ....
Aq1285	pilC2	fimbrial assembly protein PilC	28.9% ....
Aq1601	pilD	type 4 prepilin peptidase	34.8% ....
Aq745	pilT	twitching motility protein PilT	51.4% ....
Aq2151	pilU	twitching mobility protein	41.6% ....
Aq1870	secA	preprotein translocase SecA subunit	44.9% ....
Aq973	secD	protein export membrane protein SecD	36.0% ....
Aq1602	secF	protein-export membrane protein	41.4% ....
Aq079	secY	preprotein translocase SecY	44.2% ....
Aq2080	sppA	proteinase IV	43.4% ....
Aq1971	tapB	type IV pilus assembly protein TapB	42.2% ....
Aq1340	tig	trigger factor	27.4% ....
Central Intermediary Metabolism			
One-carbon metabolism			
Aq1429	metF	5,10-methylenetetrahydrofolate reductase	43.3% ....
Aq1154	metK	S-adenosylmethionine synthetase	49.2% ....
Aq1180	sahH	S-adenosylhomocysteine hydrolase	60.9% ....
Cytoplasmic polysaccharides			
Aq1407	bcsA	cellulose synthase catalytic subunit	39.5% ....
Aq1401	celY	endoglucanase fragment	33.0% ..
Aq721	glgA	glycogen synthase	38.1% ....
Aq722	glgB	1,4-alpha-glucan branching enzyme	56.5% ....
Aq717	glgP	glycogen phosphorylase	37.0% ....
Aq723	malM	4-alpha-glucanotransferase (amylomaltase)	43.4% ....
Tri-carboxylic acid cycle			
Aq1784	aco	aconitase	36.1% ...
Aq1195	forA1	ferredoxin oxidoreductase alpha subunit	31.5% ...
Aq1167	forA2	ferredoxin oxidoreductase alpha subunit	32.3% ....
Aq1196	forB1	ferredoxin oxidoreductase beta subunit	29.6% ....
Aq1168	forB2	ferredoxin oxidoreductase beta subunit	31.5% ...
Aq1200	forG1	ferredoxin oxidoreductase gamma subunit	34.5% ....
Aq1169	forG2	ferredoxin oxidoreductase gamma subunit	34.5% ...
Aq594	frdA	fumarate reductase flavoprotein subunit	51.4% ....
Aq553	frdB1	reductase iron-sulfur subunit	35.2% ....
Aq655	frdB2	fumarate reductase iron-sulfur subunit	35.1% ....
Aq1780	fumB	fumarate hydratase (fumarase)	46.4% ....
Aq1679	fumX	C-terminal fumarate hydratase, class I	40.4% ....
Aq150	gltA	citrate synthase	33.0% ....
Aq1512	icd	isocitrate dehydrogenase	46.0% ....
Aq1782	mdh1	malate dehydrogenase	49.8% ....
Aq1665	mdh2	malate dehydrogenase	46.9% ...
Aq1614	oadA	oxaloacetate decarboxylase alpha chain	50.1% ....
Aq1306	sucC1	succinyl-CoA ligase beta subunit	35.1% ....
Aq1620	sucC2	succinyl-CoA ligase beta subunit	52.9% ....
Aq1888	sucD1	succinyl-CoA ligase alpha subunit	41.7% ...
Aq1622	sucD2	succinyl-CoA ligase alpha subunit	65.7% ....
Phosphate			
Aq1351	phoH	phosphate starvation-inducible protein	47.1% ....
Aq1547	ppa	inorganic pyrophosphatase	56.5% ....
Aq891	ppx	exopolyphosphatase	33.6% ....
Polyamines			
Aq728	speC	ornithine decarboxylase	30.9% ....
Aq062	speE	spermidine synthase	48.4% ....
Sulfur			
Aq1081	cysD	sulfate adenylyltransferase	46.7% ....
Aq1076	rhdA	thiosulfate sulfurtransferase	32.3% ....
Aq1799	rhdA	thiosulfate sulfurtransferase	31.7% ....
Aq455	sor	sulfur oxygenase reductase	36.7% ....
Aq1803	soxB	sulfur oxidation protein SoxB	41.3% ....
Cofactor Biosynthesis			
Lipoic acid biosynthesis			
Aq1355	lipA	Lipoic acid synthetase	48.9% ...
Biotin			
Aq170	bioA	DAPA aminotransferase	51.7% ....
Aq975	bioB	biotin synthetase	42.0% ....
Aq557	bioD	dethiobiotin synthetase	41.5% ....
Aq626	bioF	8-amino-7-oxononanoate synthase	45.1% ....
Aq1659	bioW	6-carboxyhexanoate-CoA ligase (pimeloyl CoA synthase)	47.3% ....
Aq566	birA	biotin [acetyl-CoA-carboxylase] ligase	37.5% ....
Folic acid			
Aq2045	folC	folylpolyglutamate synthetase	31.8% ....
Aq1898	folD	methylenetetrahydrofolate dehydrogenase	53.2% ....
Aq239	folE	GTP cyclohydrolase I	57.1% ....
Aq162	folK	folate biosynthesis 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase	43.7% ....
Aq1468	folP	dihydropteroate synthase	45.8% ....
Aq1144	pabB	p-aminobenzoate synthetase	41.5% ...
Aq1606	pabC	aminodeoxychorismate lyase	29.0% ....
Heme			
Aq207	cobA	uroporphyrin-III c-methyltransferase	52.1% ....
Aq1237	cysG	siroheme synthase	36.9% ....
Aq334	dcuP	uroporphyrinogen decarboxylase	41.4% ....
Aq816	gsa	glutamate-1-semialdehyde aminotransferase	56.5% ....
Aq1279	hemA	glutamyl tRNA reductase (delta-aminolevulinate synthase)	38.7% ....
Aq2109	hemB	porphobilinogen synthase	64.5% ....
Aq263	hemC	porphobilinogen deaminase	53.1% ....
Aq1424	hemF	oxygen-independent coproporphyrinogen III oxidase	33.1% ....
Aq2015	hemG	protoporphyrinogen oxidase	30.3% ...
Aq948	hemH	ferrochelatase	46.4% ....
Aq099	hemK	protoporphyrinogen oxidase	32.2% ....
Aq2124	hemN	oxygen-independent coproporphyrinogen II	50.2% ....
Molybdopterin			
Aq2183	moaA2	molybdenum cofactor biosynthesis protein A	47.0% ....
Aq527	moaC	molybdenum cofactor biosynthesis moaC	45.0% ..
Aq2181	moaE	molybdopterin converting factor subunit 2	39.3% ...
Aq1326	mobB	molybdopterin-guainine dinucleotide biosynthesis protein B	44.4% .
Aq030	moeA1	molybdenum cofactor biosynthesis protein A	36.8% ....
Aq1329	moeB	molybdopterin biosynthesis protein MoeB	54.1% ....
Aq061	mog	molybdenum cofactor biosynthesis MOG	55.5% ....
Aq049	phhB	pterin-4a-carbinolamine dehydratase	37.9% ....
Panthenate			
Aq815	dfp	pantothenate metabolism flavoprotein	41.2% ....
Aq1973	panB	3-methyl-2-oxobutanoate hydroxymethyltransferase	45.5% ....
Aq2132	panC	pantothenate synthetase	47.4% ....
Aq476	panD	aspartate 1-decarboxylase	46.0% ....
Pyridine nucleotides			
Aq1889	nadA	quinolinate synthetase A	44.3% ....
Aq777	nadB	L-aspartate oxidase	36.7% ...
Aq869	nadC	quinolinate phosphoribosyl transferase	47.0% ....
Aq959	nadE	NH(3)-dependent NAD+ synthetase	39.6% ....
Pyridoxal phosphate			
Aq852	pdxA	pyridoxal phosphate biosynthetic protein PdxA	36.8% ....
Aq1423	pdxJ	pyridoxal phosphate synthetase	88.2% ....
Quinones			
Aq895	ispB	octoprenyl-diphosphate synthase	35.7% ....
Aq052	ubiA	4-hydroxybenzoate octaprenyltransferase	41.4% ....
Riboflavin			
Aq350	ribA	GTP cyclohydrolase II	61.7% ....
Aq1707	ribC	riboflavin synthase alpha chain	45.3% ....
Aq138	ribD1	riboflavin specific deaminase	46.0% ....
Aq436	ribD2	riboflavin specific deaminase	42.9% ....
Aq139	ribF	riboflavin kinase	38.4% ....
Aq132	ribH	riboflavin synthase beta subunit	51.0% ....
Thiamine			
Aq1204	thiC	thiamine biosynthesis protein	67.1% ....
Aq1960	thiD	HMP-P kinase	40.5% ....
Aq1366	thiE1	thiamine phosphate synthase	36.3% ....
Aq558	thiE2	thiamine phosphate synthase	39.5% ....
Aq2178	thiG	thiamine biosynthesis, thiazole moiety	52.5% ....
Aq2119	thiL	thiamine monophosphate kinase	34.5% ....
Thio- and glutaredoxin			
Aq443	gua	glutaredoxin-like protein	33.8% ....
Aq1916	trxA1	thioredoxin	58.9% ...
Aq1811	trxA2	thioredoxin	32.2% ..
Aq500	trxB	thioredoxin reductase	39.8% ....
Energy Metabolism			
Aq1342	gph	phosphoglycolate phosphatase	33.9% ....
ATP-Proton Motive Force			
Aq679	atpA	ATP synthase F1 alpha subunit	64.3% ....
Aq179	atpB	ATP synthase F0 subunit a	36.4% ....
Aq673	atpC	ATP synthase F1 epsilon subunit	37.4% ....
Aq2038	atpD	ATP synthase F1 beta subunit	67.4% ....
Aq177	atpE	ATP synthase F0 subunit c	53.8% ...
Aq1586	atpF1	ATP synthase F0 subunit b	26.3% ....
Aq1587	atpF2	ATP synthase F0 subunit b	25.5% ....
Aq2041	atpG	ATP synthase F1 gamma subunit	39.9% ....
Aq1588	atpH	ATP synthase F1 delta chain	28.1% ....
Dehydrogenases			
Aq1362	adh1	alcohol dehydrogenase	35.4% ....
Aq1240	adh2	alcohol dehydrogenase	28.8% ....
Aq186	aldH1	aldehyde dehydrogenase	41.9% ....
Aq227	aldH2	aldehyde dehydrogenase	28.0% ....
Aq1145	dhaT	1,3 propanediol dehydrogenase	36.6% ....
Aq232	dhsU	flavocytochrome C sulfide dehydrogenase	33.6% ....
Aq1769	dld1	D-lactate dehydrogenase	45.3% ....
Aq1234	dmsA	DMSO reductase chain A	25.0% ....
Aq1232	dmsB	DMSO reductase chain B	38.4% ...
Aq1231	dmsC	DMSO reductase chain C	29.5% .
Aq1051	fdhE	formate dehydrogenase formation protein FdhE	25.9% ....
Aq1039	fdoG	formate dehydrogenase alpha subunit	50.0% ....
Aq1046	fdoH	formate dehydrogenase beta subunit	45.7% ....
Aq1049	fdoI	formate dehydrogenase gamma subunit	38.4% ....
Aq1903	gcsP1	glycine dehydrogenase (decarboxylating)	49.6% ....
Aq1109	gcsP2	glycine dehydrogenase (decarboxylating)	46.8% ....
Aq1639	glpC	oxido/reductase iron sulfur protein	27.1% ....
Aq395	hdrA	heterodisulfide reductase subunit A	39.7% ....
Aq400	hdrB	heterodisulfide reductase subunit B	32.5% ....
Aq398	hdrC	heterodisulfide reductase subunit C	35.7% .
Aq961	hdrD	heterodisulfide reductase	29.5% ....
Aq038	hibD	3-hydroxyisobutyrate dehydrogenase	34.6% ....
Aq727	ldhA	D-lactate dehydrogenase	33.5% ....
Aq736	lpdA	dihydrolipoamide dehydrogenase	37.0% ....
Aq217	narB	nitrate reductase narB	39.1% ....
Aq206	nirB	nitrite reductase (NAD(P)H) large subunit	35.3% ...
Aq835	nox	NADH oxidase	33.1% ....
Aq024	nsd	nucleotide sugar dehydrogenase	47.0% ....
Aq135	nueM	NADH dehydrogenase (ubiquinone)	28.2% ....
Aq1010	udh	dehydrogenase	29.7% ....
Electron transport			
Aq2191	coxA1	cytochrome c oxidase subunit I	42.4% ....
Aq2192	coxA2	cytochrome c oxidase subunit I	38.1% ....
Aq2190	coxB	cytochrome c oxidase subunit II	27.4% ....
Aq2188	coxC	cytochrome c oxidase subunit III	28.6% ....
Aq153	ctaA	heme O oxygenase	28.1% ....
Aq042	cyc	cytochrome c	25.8% ...
Aq792	cycB1	cytochrome c552	29.9% ..
Aq1550	cycB2	cytochrome C552	38.7% ....
Aq1357	cydA	cytochrome oxidase d subunit I	38.8% ....
Aq1358	cydB	cytochrome oxidase d subunit II	31.2% ....
Aq067	dmsB	dimethylsulfoxide reductase chain B	40.2% ....
Aq235	fccB'	sulfide dehydrogenase, flavoprotein subunit	38.0% ...
Aq919a	fdx1	ferredoxin	37.1% ...
Aq1171a	fdx2	ferredoxin	43.9% ..
Aq1192a	fdx3	ferredoxin	35.0% ...
Aq108a	fdx4	ferredoxin	56.6% ...
Aq211	fhp	flavohemoprotein	43.4% ....
Aq2096	fldX	flavodoxin	32.5% ....
Aq045	petA	Rieske-1 iron sulfur protein	34.3% ....
Aq044	petB	cytochrome b	38.3% ...
Aq234	soxF	Rieske-1 iron sulfur protein	29.0% ....
Aq2186	sqr	sulfide-quinone reductase	41.0% ....
Glycolysis and gluconeogenesis			
Aq185	eno	enolase	65.0% ....
Aq1390	fba	fructose-1,6-bisphosphate aldolase class II	39.9% ....
Aq1065	gap	glyceraldehyde-3-phosphate dehydrogenase	59.5% ....
Aq434	glpK	glycerol kinase	51.0% ....
Aq1744	gpmA	phosphoglycerate mutase	27.9% ....
Aq1634	gspA	glycerol-3-phosphate dehydrogenase (NAD+)	40.5% ....

Aq1708	pfkA	phosphofructokinase	49.4% ....
Aq750	pgi	glucose-6-phosphate isomerase	37.8% ....
Aq118	pgk	phosphoglycerate kinase	54.5% ....
Aq1990	pgmA	phosphoglycerate mutase	33.2% ....
Aq501	pmu	phosphoglucomutase/phosphomannomutase	37.8% ....
Aq2142	ppsA	phosphoenolpyruvate synthase	56.3% ....
Aq1520	pycA	pyruvate carboxylase c-terminal domain	46.6% ....
Aq1517	pycB	pyruvate carboxylase n-terminal domain	57.1% ....
Aq360	timA	triose phophate isomerase	52.2% ....
Hydrogenase			
Aq665	hoxZ	Ni/Fe hydrogenase B-type cytochrome subunit	40.4% ....
Aq667	hupD	HupD hydrogenase related function	40.9% ....
Aq666	hupE	HupE hydrogenase related function	38.3% ....
Aq1021	hypA	hydrogenase accessory protein HypA	39.8% ....
Aq671	hypB	hydrogenase expression/formation protein B	50.6% ....
Aq1157	hypD	hydrogenase expression/formation protein HypD	56.1% ....
Aq662	mbhL1	hydrogenase large subunit	50.6% ....
Aq960	mbhL2	hydrogenase large subunit	44.3% ....
Aq804	mbhL3	hydrogenase large subunit	27.9% ....
Aq660	mbhS1	hydrogenase small subunit	66.6% ....
Aq965	mbhS2	hydrogenase small subunit	51.3% ....
Aq802	mbhS3	hydrogenase small subunit	36.7% ....
Aq1591	shyS	soluble hydrogenase small subunit	41.6% ....
Sugar metabolism			
Aq968	cbbE2	ribulose-5-phosphate 3-epimerase	47.2% ....
Aq1658	fucA1	fuculose-1-phosphate aldolase	31.8% ....
Aq1979	fucA2	fuculose-1-phosphate aldolase	29.7% ....
Aq498	gnd	6-phosphogluconate dehydrogenase	45.2% ....
Aq497	gsdA	glucose-6-phosphate 1-dehydrogenase	32.3% ....
Aq1138	rpiB	ribose 5-phosphate isomerase B	54.5% ....
Aq119	talC	transaldolase	71.1% ....
Aq1765	tktA	transketolase	52.4% ....
NADH dehydrogenase			
Aq1385	nuoA1	NADH dehydrogenase I chain A	42.0% ....
Aq1310	nuoA2	NADH dehydrogenase I chain A	44.9% ....
Aq1312	nuoB	NADH dehydrogenase I chain B	60.1% ...
Aq551	nuoD1	NADH dehydrogenase I chain D	37.7% ....
Aq1314	nuoD2	NADH dehydrogenase I chain D	42.2% ....
Aq574	nuoE	NADH dehydrogenase I chain E	36.8% ....
Aq573	nuoF	NADH dehydrogenase I chain F	20.5% ..
Aq437	nuoG	NADH dehydrogenase I chain G	35.4% ..
Aq1315	nuoH1	NADH dehydrogenase I chain H	41.0% ....
Aq1373	nuoH2	NADH dehydrogenase I chain H	42.1% ....
Aq1374	nuoH3	NADH dehydrogenase I chain H	38.9% ....
Aq1317	nuoI1	NADH dehydrogenase I chain I	30.5% ...
Aq1375	nuoI2	NADH dehydrogenase I chain I	29.2% ...
Aq1318	nuoJ1	NADH dehydrogenase I chain J	35.4% ....
Aq1377	nuoJ2	NADH dehydrogenase I chain J	30.6% ....
Aq1319	nuoK1	NADH dehydrogenase I chain K	51.1% ....
Aq1378	nuoK2	NADH dehydrogenase I chain K	48.4% ....
Aq1320	nuoL1	NADH dehydrogenase I chain L	39.0% ....
Aq866	nuoL2	NADH dehydrogenase I chain L	30.2% ...
Aq1379	nuoL3	NADH dehydrogenase I chain L	43.1% ....
Aq1321	nuoM1	NADH dehydrogenase I chain M	43.6% ....
Aq1382	nuoM2	NADH dehydrogenase I chain M	36.9% ....
Aq1322	nuoN1	NADH dehydrogenase I chain N	34.1% ....
Aq1383	nuoN2	NADH dehydrogenase I chain N	32.8% ....
Lipid metabolism			
Aq2058	aas	2-acylglycerophosphoethanolamine acyltransferase	37.1% ....
Aq1206	accA	acetyl-CoA carboxylase alpha subunit	57.1% ...
Aq1363	accB	biotin carboxyl carrier protein	44.6% ....
Aq1664	accC1	biotin carboxylase	54.4% ....
Aq1470	accC2	biotin carboxylase	56.5% ....
Aq445	accD	acetyl-CoA carboxyltransferase beta subunit	56.9% ....
Aq1717a	acpP	acyl carrier protein	71.2% ....
Aq813	acpS	holo-[acyl-carrier protein] synthase	30.8% ....
Aq2104	acs	acetyl-coenzyme A synthetase	54.0% ....
Aq2103	acs'	acetyl-coenzyme A synthetase c-terminal fragment	61.2% ....
Aq1249	cds	phosphatidate cytidylyltransferase	29.2% ....
Aq1737	cfa	cyclopropane-fatty-acyl-phospholipid synthase	37.5% ....
Aq892	fabD	malonyl-CoA:Acyl carrier protein transacylase	42.1% ....
Aq1717	fabF	3-oxoacyl-[acyl-carrier-protein] synthase II	58.4% ....
Aq1716	fabG	3-oxoacyl-[acyl-carrier-protein] reductase	52.9% ....
Aq1099	fabH	3-oxoacyl-[acyl-carrier-protein] synthase III	47.0% ....
Aq1552	fabI	enoyl-[acyl-carrier-protein] reductase (NADH)	49.6% ....
Aq056	fabZ	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase	58.7% ....
Aq999	fadD	long-chain-fatty-acid CoA ligase	30.0% ...
Aq1638	lplA	lipoate-protein ligase A	28.1% ..
Aq958	pgsA	phosphotidylglycerophosphate synthase	37.3% ..
Aq2154	pgsA	phosphotidylglycerophosphate synthase	38.9% ...
Aq1101	plsX	PlsX protein	43.7% ....
Purines, Pyrimidines, Nucleotides and Nucleosides			
Aq094	nrdA	ribonucleotide reductase alpha chain	35.0% ....
Aq1505	nrdF	ribonucleotide reductase beta chain	36.2% .
Purines			
Aq568	deoD	purine nucleoside phosphorylase	33.1% ....
Aq236	guaA	GMP synthase	58.4% ....
Aq2023	guaB	inosine monophosphate dehydrogenase	65.4% ....
Aq544	hpt	hypoxanthine-guanine phosphoribosyltransferase	48.2% ....
Aq078	kad	adenylate kinase	50.0% ....
Aq1590	ndk	nucleoside diphosphate kinase	48.2% ....
Aq1636	prs	phosphoribosylpyrophosphate synthetase	55.2% ....
Aq1290	purA	adenylosuccinate synthetase	49.2% ....
Aq597	purB	adenylosuccinate lyase	52.4% ....
Aq2117	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase	52.5% ....
Aq742	purD	phosphoribosylamine-glycine ligase	54.2% ....
Aq1178	purE	phosphoribosylaminoimidazole carboxylase	64.6% ....
Aq1175	purF	amidophosphoribosyltransferase	42.7% ....
Aq1963	purH	phosphoribosylaminoimidazolecarboxamide formyltransferase	48.2% ....
Aq245	purK	phosphoribosyl aminoimidazole carboxylase	35.6% ....
Aq1836	purL	phosphoribosylformylglycinamidine synthase II	49.3% ...
Aq769	purM	phosphoribosylformylglycinamidine cyclo-ligase	50.0% ....
Aq857	purN	phosphoribosylglycinamide formyltransferase	48.3% ....
Aq1105	purQ	phosphoribosyl formylglycinamidine synthase I	51.1% ....
Aq1818	purU	formyltetrahydrofolate deformylase	56.3% ....
Pyrimidines			
Aq410	carA	carbamoyl phosphate synthetase small subunit	52.2% ....
Aq1172	carB	carbamoyl-phosphate synthase large subunit	60.7% ....
Aq2101	carB	carbamoyl-phosphate synthase, large subunit	63.1% ....
Aq2153	cmk	cytidylate kinase	38.5% ....
Aq1607	dcd	deoxycytidine triphosphate deaminase	39.5% ....
Aq220	dut	deoxyuridine 5'triphosphate nucleotidohydrolase	
Aq409	pyrB	aspartate carbamoyltransferase catalytic chain	42.0% ....
Aq806	pyrC	dihydroorotase	37.3% ...
Aq046	pyrD	dihydroorotase dehydrogenase	50.5% ....
Aq1305	pyrDB	dihydroorotate dehydrogenase electron transfer subunit	34.7% ....
Aq1580	pyrF	orotidine-5'-phosphate decarboxylase	37.2% ....
Aq1334	pyrG	CTP synthetase	57.5% ....
Aq713	pyrH	UMP kinase	62.1% ....
Aq640	thy	thymidylate synthase complementing protein	30.5% ...
Aq969	tmk	thymidylate kinase	35.1% ....
Aq1907	umpS	uridine 5-monophosphate synthase	42.1% ....
Aq2163	uraP	uracil phosphoribosyltransferase	42.0% ....
Regulation			
Aq1058	acrR1	transcriptional regulatcr (TetR/AcrR family)	34.1% ...
Aq2179	acrR2	transcriptional regulator (TetR/AcrR family)	31.0% ...
Aq281	acrR3	transcriptional regulator (TetR/AcrR family)	29.7% ....
Aq1387	arsR	transcriptional regulator (ArsR family)	35.3% ....
Aq1724	degT	transcriptional regulator (DegT/Dnrl/Eryc1 family)	34.1% ....
Aq534	draG	ADP-ribosylglycohydrolase	32.1% ....
Aq831	exsB	trans-regulatory protein ExsB	38.5% ....
Aq490	fnr	transcriptional regulator (Crp/Fnr family)	29.5% ....
Aq1207	furR1	transcriptional regulator (FurR family)	37.9% ....
Aq1418	furR2	transcriptional regulator (FurR family)	34.6% ....
Aq213	glnBi	PII-like protein GlnBi	48.0% ..
Aq1908	hflX	GTP-binding protein HflX	40.3% ....
Aq1115	hksP1	histidine kinase sensor protein	27.7% ..
Aq316	hksP2	histidine kinase sensor protein	28.1% ....
Aq905	hksP3	histidine kinase sensor protein	23.6% ...
Aq231	hksP4	histidine kinase sensor protein	28.2% ....
Aq1156	hoxX	hydrogenase regulation HoxX	46.7% ....
Aq093	hth	transcriptional regulator (H-T-H)	50.2% ....
Aq1019	hypE	hydrogenase expression/formation protein	44.3% ....
Aq672	hypF	transcriptional regulatory protein HypF	44.8% ....
Aq764	iclR	transcriptional regulator (IclR family)	30.4% ....
Aq638	lysR1	transcriptional regulator (LysR family)	32.8% ....
Aq1038	lysR2	transcriptional regulator (LysR family)	28.9% ....
Aq702	merR	transcriptional regulator (MerR family)	32.8% ....
Aq218	nifA	transcriptional regulator (NifA family)	42.8% ....
Aq1117	ntrC1	transcriptional regulator (NtrC family)	41.0% ....
Aq1792	ntrC2	transcriptional regulator (NtrC family)	40.2% ....
Aq230	ntrC3	transcriptional regulator (NtrC family)	40.0% ....
Aq164	ntrC4	transcriptional regulator (NtrC family)	38.3% ....
Aq2069	obg	GTP-binding protein	54.9% ..
Aq319	phoB	transcriptional regulator (PhoB-like)	41.6% ....
Aq906	phoU	transcriptional regulator (PhoU-like)	41.9% ....
Aq844	spoT	(p)ppGpp 3-pyrophosphohydrolase	47.2% ...
Aq1496	xylR	transcriptional regulator (NagC/XylR family)	29.3% ....
DNA Replication and Repair			
Aq358	dinG	ATP-dependent helicase (DinG family)	27.9% ....
Aq322	dnaA	chromosome replication initiator protein DnaA	36.5% ....
Aq1472	dnaB	replicative DNA helicase	40.3% ....
Aq910	dnaC	DNA replication protein DnaC	26.4% ....
Aq1008	dnaE	DNA polymerase III alpha subunit	41.9% ...
Aq1493	dnaG	DNA primase	39.8% ...
Aq1882	dnaN	DNA polymerase III beta chain	32.1% ....
Aq932	dnaQ	DNA polymerase III epsilon subunit	40.0% ....
Aq1855	dnaX	DNA polymerase III gamma subunit	36.6% ....
Aq1422	dpbF	DNA polymerase beta family	39.1% ....
Aq1693	dplF	N-terminus of phage SPO1 DNA polymerase	37.3% ....
Aq980	gyrA	DNA gyrase A subunit	43.6% ....
Aq1026	gyrB	gyrase B	55.2% ..
Aq2057	helX	DNA helicase	49.7% ....
Aq1484a	himA	DNA binding protein HU	40.2% ....
Aq2174	ihfB	integration host factor beta subunit	35.8% ....
Aq1394	lig	DNA ligase (ATP dependent)	50.8% ....
Aq633	ligA	DNA ligase (NAD dependent)	45.7% ....
Aq1578	mutL	DNA mismatch repair protein MutL	72.3% ....
Aq308	mutS1	DNA mismatch repair protein MutS	77.5% ....
Aq1242	mutS2	DNA mismatch repair protein MutS	37.0% ....
Aq1449	mutT	8-OXO-dGTPase domain (mutT domain)	46.3% ....
Aq282	mutY1	endonuclease III	53.6% ....
Aq172	mutY2	endonuclease III	51.8% ....
Aq496	mutY3	endonuclease III	43.4% ....
Aq1629	nfo	deoxyribonuclease IV	39.0% ....
Aq710	nucl	thermococcal nuclease homolog	36.4% ....
Aq1495	ogt	O-6-methylguanine-DNA-alkyltransferase	36.9% ...
Aq1628	pol	DNA polymerase I 3'-5' exo domain	43.2% ....
Aq1967	polA	DNA polymerase I (PolI)	30.5% ....
Aq1610	radC	DNA repair protein RadC	39.0% ....
Aq2150	recA	recombination protein RecA	88.5% ....
Aq2053	recG	ATP-dependent DNA helicase RecG	38.9% ....
Aq2155	recJ	single-strand-DNA-specific exonuclease RecJ	31.8% ....
Aq561	recN	recombination protein RecN	27.7% ....
Aq1478	recR	recombination protein RecR	38.3% ....
Aq793	rep	ATP-dependent DNA helicase REP	33.4% ....
Aq1886	sbcD	ATP-dependent dsDNA exonuclease	29.9% ....
Aq064	ssb	single stranded DNA-binding protein	39.4% ...
Aq657	topA	topoisomerase I	39.6% ....
Aq1159	topG1	reverse gyrase	41.6% ....
Aq886	topG2	reverse gyrase	35.1% ....
Aq686	uvrA	repair excision nuclease subunit A	61.0% ....
Aq1856	uvrB	repair excision nuclease subunit B	53.9% ....
Aq2126	uvrC	repair excision nuclease subunit C	32.5% ....
Transcription			
RNA polymerase and transcription factors			
Aq613	deaD	ATP-dependent RNA helicase DeaD	42.3% ....
Aq357a	flgM	anti sigma factor FlgM	20.6% ....
Aq1218	fliA	RNA polymerase sigma factorFliA	37.2% ....
Aq259	nusA	transcription termination NusA	45.4% ...
Aq133	nusB	transcription termination NusB	32.3% ....
Aq1931	nusG	transcription antitermination protein NusG	46.3% ..
Aq873	rho	transcriptional terminator Rho	59.6% ....
Aq070	rpoA	RNA polymerase alpha subunit	40.4% ....
Aq1939	rpoB	RNA polymerase beta subunit	44.0% ....
Aq1945	rpoC	RNA polymerase beta prime subunit	46.9% ....
Aq1490	rpoD	RNA polymerase sigma factor RpoD	41.6% ....
Aq599	rpoN	RNA polymerase sigma factor RpoN	30.6% ....
Aq1452	rpoS	RNA polymerase sigma factor RpoS	40.5% ....
RNA modification			
Aq1816	ksgA	dimethyladenosine transferase	36.1% ....
Aq1067	miaA	tRNA delta-2-isopentenylpyrophosphate (IPP) transferase	38.2% ....
Aq411	pcnB1	poly A polymerase	28.5% ....
Aq2158	pcnB2	poly A polymerase	33.9% ..
Aq221	phpA	polyribonucleotide nucleotidyltransferase	45.0% ....
Aq894	queA	queuosine biosynthesis protein	46.9% ....
Aq946	rnc	RNase III	35.8% ....
Aq1955	rnhB	RNase HII	48.4% ....
Aq924	rnpH	RNase PH	64.0% ....
Aq1661	spoU	rRNA methylase SpoU	44.0% ....
Aq1308	tgt	queuine tRNA-ribosyltransferase	52.6% ....
Aq841	trm1	N2,N2-dimethylguanosine tRNA	

		methyltransferase	34.6% ....
Aq1489	trmD	tRNA guanine-N1 methyltransferase	42.9% ....
Aq749	truA	pseudouridine synthase I	33.1% ....
Aq705	truB	tRNA pseudouridine 55 synthase	38.2% ....
Aq1890	tsnR	rRNA methylase	36.4% ....
Aq2046	vacB	VacB protein (ribonuclease II family)	37.9% ....
Aq257	ygcA	RNA methyltransferase (TrmA-family)	28.8% ....
Translation			
Aq2131	fmt	methionyl-tRNA formyltransferase	45.7% ....
Aq247	gatA	glutamyl-tRNA(Gln) amidotransferase subunit A	53.6% ....
Aq461	gatB	glutamyl-tRNA(Gln) amidotransferase subunit B	48.8% ....
Aq2147a	gatC	glutamyl-tRNA(Gln) amidotransferase subunit C	41.1% ....
Aq346	pth	peptidyl-tRNA hydrolase	48.8% ....
Aminoacyl tRNA synthetases			
Aq1293	alaS	alanyl-tRNA synthetase	46.6% ....
Aq923	argS	arginyl-tRNA synthetase	39.4% ....
Aq1677	aspS	aspartyl-tRNA synthetase	51.3% ....
Aq1068	cysS	cysteinyl-tRNA synthetase	45.0% ....
Aq763	genX	lysyl-tRNA synthetase (genX) homolog	38.6% ....
Aq1221	gltX	glutamyl-tRNA synthetase	48.5% ....
Aq945	glyQ	glycyl-tRNA synthetase alpha subunit	61.9% ....
Aq2141	glyS	glycyl-tRNA synthetase beta subunit	37.1% ....
Aq122	hisS1	histidyl-tRNA synthetase	43.3% ....
Aq1155	hisS2	histidyl-tRNA synthetase	34.9% ....
Aq305	ileS	isoleucyl-tRNA synthetase	82.1% ..
Aq351	leuS	leucyl-tRNA synthetase alpha subunit	50.7% ....
Aq1770	leuS'	leucyl-tRNA synthetase beta subunit	47.2% ....
Aq1202	lysU	lysyl-tRNA synthetase	53.2% ...
Aq1257	metG	methionyl-tRNA synthetase alpha subunit	45.0% ....
Aq422	metG'	methionyl-tRNA synthetase beta subunit	64.2% ....
Aq953	pheS	phenylalanyl-tRNA synthetase alpha subunit	51.9% ....
Aq1730	pheT	phenylalanyl-tRNA synthetase beta subunit	35.4% ....
Aq365	proS	proline-tRNA synthetase	44.1% ....
Aq298	serS	seryl-tRNA synthetase	59.4% ....
Aq1667	thrS	threonyl-tRNA synthetase	48.5% ....
Aq992	trpS	tryptophanyl-tRNA synthetase	38.4% ....
Aq1751	tyrS	tyrosyl tRNA synthetase	56.2% ....
Aq1413	valS	valyl-tRNA synthetase	33.2% ....
Ribosomal Proteins			
Aq1935	rplA	ribosomal protein L01	57.9% ....
Aq013	rplB	ribosomal protein L02	46.9% ....
Aq009	rplC	ribosomal protein L03	53.8% ....
Aq011	rplD	ribosomal protein L04	51.3% ....
Aq1652	rplE	ribosomal protein L05	67.0% ....
Aq1649	rplF	ribosomal protein L06	46.2% ....
Aq2042	rplI	ribosomal protein L09	35.6% ....
Aq1936	rplJ	ribosomal protein L10	36.5% ....
Aq1933	rplK	ribosomal protein L11	71.4% ....
Aq1937	rplL	ribosomal protein L7/L12	75.4% ....
Aq1877	rplM	ribosomal protein L13	60.6% ....
Aq1654	rplN	ribosomal protein L14	59.5% ....
Aq1642	rplO	ribosomal protein L15	57.4% ....
Aq018	rplP	ribosomal protein L16	59.3% ....
Aq069	rplQ	ribosomal protein L17	48.7% ....
Aq1648	rplR	ribosomal protein L18	62.7% ....
Aq1954	rplS	ribosomal protein L19	59.8% ...
Aq952	rplT	ribosomal protein L20	63.5% ....
Aq016a	rplV	ribosomal protein L22	47.3% ....
Aq012	rplW	ribosomal protein L23	52.2% ....
Aq1653	rplX	ribosomal protein L24	50.8% ....
Aq1644	rpmD	ribosomal protein L30	46.4% ..
Aq1930a	rpmG	ribosomal protein L33	67.9% ....
Aq792a	rpmI	ribosomal protein L35	48.3% ....
Aq1485	rpsA	ribosomal protein S01	32.6% ....
Aq2007	rpsB	ribosomal protein S02	60.3% ....
Aq017	rpsC	ribosomal protein S03	54.0% ....
Aq072	rpsD	ribosomal protein S04	51.9% ....
Aq1645	rpsE	ribosomal protein S05	60.6% ...
Aq063	rpsF	ribosomal protein S06	32.7% ....
Aq1832	rpsG1	ribosomal protein S07	52.5% ....
Aq734	rpsG2	ribosomal protein S07	51.9% ....
Aq1651	rpsH	ribosomal protein S08	39.9% ....
Aq1878	rpsI	ribosomal protein S09	50.5% ...
Aq008	rpsJ	ribosomal protein S10	55.9% ....
Aq073	rpsK	ribosomal protein S11	60.7% ....
Aq735	rpsL1	ribosomal protein S12	78.9% ....
Aq1834	rpsL2	ribosomal protein S12	78.9% ....
Aq074	rpsM	ribosomal protein S13	61.9% ....
Aq1651a	rpsN	ribosomal protein S14	51.6% ....
Aq226a	rpsO	ribosomal protein S15	61.6% ....
Aq123	rpsP	ribosomal protein S16	36.6% ...
Aq020	rpsQ	ribosomal protein S17	59.6% ....
Aq064a	rpsR	ribosomal protein S18	48.5% ....
Aq015	rpsS	ribosomal protein S19	63.1% ..
Aq1767	rpsT	ribosomal protein S20	40.0% ...
Aq867a	rpsU	ribosomal protein S21	38.2% ....
Translation factors			
Aq1364	efp	elongation factor P	48.6% ....
Aq2114	eif	initiation factor eIF-2B alpha subunit	58.4% ...
Aq712	frr	ribosome recycling factor	43.0% ....
Aq001	fusA	elongation factor EF-G	91.9% ....
Aq075a	infA	initiation factor IF-1	69.1% ....
Aq2032	infB	initiation factor IF-2	48.5% ....
Aq1777	infC	initiation factor IF-3	53.6% ....
Aq876	prfA	peptide chain release factor RF-1	54.8% ....
Aq1840	prfB	peptide chain release factor RF-2	49.9% ....
Aq1033	selB	elongation factor SelB	30.4% ....
Aq715	tsf	elongation factor EF-Ts	35.8% ....
Aq005	tufA1	elongation factor EF-Tu	74.4% ....
Aq1928	tufA2	elongation Factor EF-Tu	73.9% ....
Protein modification			
Aq731	ccdA	cytochrome c-type biogenesis protein	32.0% ....
Aq579	def	polypeptide deformylase	41.4% ....
Aq2093	dsbC	thiol:disulfide interchange protein	27.6% ....
Aq055	hemX1	cytochrome c biogenesis protein	26.2% ....
Aq2043	hemX2	cytochrome c biogenesis protein	36.2% ....
Aq1053	nifS1	FeS cluster formation protein NifS	38.5% ....
Aq739	nifS2	FeS cluster formation protein NifS	45.5% ....
Aq1871	pmbA	peptide maturation	25.6% ....
Aq2102	prmA	ribosomal protein L11 methyltransferase	35.1% ....
Aq567	rimI	ribosomal-protein-alanine acetyltransferase	37.9% ....
Aq576	stpK	ser/thr protein kinase	30.8% ....
Aq152	tlpA	thiol disulfide interchange protein	37.6% ...
Proteases			
Aq1950	aprV	serine protease	26.5% ...
Aq1672	clpB	ATPase subunit of ATP-dependent protease	46.8% ....
Aq1296	clpC	ATP-dependent Clp protease	54.9% ....
Aq1339	clpP	ATP-dependent Clp protease proteolytic subunit	65.4% ....
Aq1337	clpX	ATP-dependent protease ATPase subunit clpX	66.1% ....
Aq1015	col	collagenase	41.3% ....
Aq801	gcp	sialoglycoprotease	45.5% ....
Aq1671	hslV	heat shock protein HsLV	57.6% ....
Aq1450	htrA	periplasmic serine protease	38.3% ....
Aq242	lon	Lon protease	50.6% ....
Aq076	map	methionyl aminopeptidase	44.1% ....
Aq1459	npr	neutral protease	27.7% ..
Aq2099	pepA	leucine aminopeptidase	39.5% ....
Aq1535	pepQ	xaa-pro dipeptidase	31.9% ....
Aq618	pfpI	protease I	41.8% ....
Aq797	prc	carboxyl-terminal protease	41.8% ....
Aq552	sms	ATP-dependent protease sms	46.2% ....
Aq2204	ymxG	processing protease	28.3% ....
Transport			
Aq1222	abcT1	ABC transporter	34.7% ....
Aq620	abcT2	ABC transporter	36.8% ....
Aq1095	abcT3	ABC transporter (ABC-2 subfamily)	34.4% ....
Aq1094	abcT4	ABC transporter	37.7% ....
Aq1097	abcT5	ABC transporter (hlyB subfamily)	45.5% ....
Aq417	abcT6	ABC transporter	51.8% ....
Aq413	abcT7	ABC transporter	51.5% ....
Aq297	abcT8	ABC transporter	49.3% ....
Aq2160	abcT9	ABC transporter	45.3% ....
Aq1531	abcT10	ABC transporter	36.4% ....
Aq2122	abcT11	ABC transporter	42.5% ....
Aq2137	abcT12	ABC transporter	38.2% ....
Aq1563	abcT13	ABC transporter (MsbA subfamily)	30.5% ....
Aq695	acrD1	cation efflux system (AcrB/AcrD/AcrF family)	22.7% ....
Aq1122	acrD2	cation efflux system (AcrB/AcrD/AcrF family)	32.0% ....
Aq469	acrD3	cation efflux system (AcrB/AcrD/AcrF family)	34.2% ....
Aq786	acrD4	cation efflux (AcrB/AcrD/AcrF family)	27.7% ....
Aq112	amtB	ammonium transporter	49.0% ....
Aq682	arsA1	anion transporting ATPase	41.5% ....
Aq343	arsA2	anion transporting ATPase	33.9% ....
Aq851	corA	Mg(2+) and Co(2+) transport protein	31.1% ....
Aq724	ctrA1	cation transporting ATPase (E1-E2 family)	30.7% ....
Aq1445	ctrA2	cation transporting ATPase (E1-E2 family)	28.1% ....
Aq1125	ctrA3	cation transporting ATPase (E1-E2 family)	43.8% ....
Aq1132	czcB1	cation efflux system (czcB-like)	23.7% ....
Aq1331	czcB2	cation efflux system (czcB-like)	26.9% ....
Aq468	czcB2	cation efflux system (czcB-like)	28.5% ....
Aq1073	czcD	cation efflux system (CzcD-like)	43.4% ....
Aq911	ebs	erythrocyte band 7 homolog	50.2% ....
Aq1062	emrB	major facilitator family transporter	28.3% ....
Aq1255	feoB	ferrous iron transport protein B	32.6% ....
Aq1330	glnP	proton/sodium-glutamate symport protein	35.6% ....
Aq1268	hvsT	high affinity sulfate transporter	29.4% ....
Aq1863	kch	potassium channel protein	30.1% ...
Aq1725	lepA	G-protein LepA	59.8% ....
Aq1229	mffT	transporter (major facilitator family)	37.2% ....
Aq447	mgtC	Mg(2+) transport ATPase	36.2% ....
Aq1609	modA	molybdate periplasmic binding protein	38.2% ....
Aq086	modC	Molybdenum transport system permease	44.8% ....
Aq415	napA1	Na(+)/H(+) antiporter	27.6% ....
Aq929	napA2	Na(+)/H(+) antiporter	32.7% ....
Aq2030	napA3	Na(+)/H(+) antiporter	26.8% ....
Aq215	nasA	nitrate transporter	35.8% ...
Aq1441	oppA	transporter (extracellular solute binding protein family 5)	37.0% ....
Aq481	oppB	transporter (OppBC family)	46.2% ....
Aq1509	oppC	oligopeptide transport system permease	46.2% ....
Aq2019	pstA	phosphate transport system permease PstA	43.5% ....
Aq1055	pstB	phosphate transport ATP binding protein	68.1% ....
Aq2018	pstC	phosphate transport system permease protein C	45.2% ....
Aq2016	pstS	phosphate-binding periplasmic protein	52.4% ....
Aq2129	sbf	Na(+) dependent transporter (Sbf family)	34.9% ....
Aq098	secG	protein export membrane protein SecG	35.7% ....
Aq2077	snf	Na(+):neurotransmitter symporter (Snf family)	25.7% ....
Aq2106	ssf	Na(+):solute symportet (Ssf family)	47.4% .
Aq1988	tolQ	TolQ homolog	32.5% .
Aq1504	trkI	K+ transport protein homolog	40.6% ....
Aq031	trnS	transporter (Pho87 family)	46.8% ....
Uncategorized			
Aq1023	acuC1	acetoin utilization protein	36.9% ....
Aq2110	acuC2	acetoin utilization protein	38.6% ....
Aq158	apfA	AP4A hydrolase	36.6% ....
Aq458	bcp	bacterioferritin comigratory protein	40.6% ....
Aq542	bcpC	phosphonopyruvate decarboxylase	37.4% ....
Aq147	cobW	cobalamin synthesis related protein CobW	29.5% ...
Aq1303a	cspC	cold shock protein	67.2% ....
Aq1265	cstA	carbon starvation protein A	33.0% ....
Aq348	ctc	general stress protein Ctc	34.7% ....
Aq212	cynS	cyanate hydrolase	39.5% ....
Aq337	cysQ	CysQ protein	47.4% ....
Aq528	dedF	phenylacrylic acid decarboxylase	52.4% ....
Aq148	deoC	deoxyribose-phosphate aldolase	46.6% ....
Aq2095	dksA	dnaK suppressor protein	35.1% ....
Aq1994	era1	GTP-binding protein Era	49.7% ....
Aq1919	era2	GTP binding protein Era	43.0% ....
Aq1540	gcpE	GcpE protein	50.1% ....
Aq1052	gcsH1	glycine cleavage system protein H	28.6% ....
Aq1657	gcsH2	glycine cleavage system protein H	39.8% ...
Aq944	gcsH3	glycine cleavage system protein H	36.7% ...
Aq1108	gcsH4	glycine cleavage system protein H	44.8% ...
Aq1458	gcvT	aminomethyltransferase (glycine cleavage system T protein)	42.2% ....
Aq108b	hfq	host factor I	53.5% ....
Aq101	hly	hemolysin	33.7% ...
Aq2120	hlyC	hemolysin homolog protein	29.3% ....
Aq1091	hyIA	hemolysin	33.5% ...
Aq708	hyuA	N-methylhydantoinase A	39.8% ....
Aq1925	hyuB	N-methylhydantoinase B	43.1% ....
Aq1579	iagB	invasion protein IagB	38.3% ...
Aq1983	imp2	myo-inositol-1(or 4)-monophosphatase	36.0% ....
Aq748	ispA	geranylgeranyl pyrophosphate synthase	40.7% ....
Aq1739	lytB	LytB protein	43.9% ....
Aq1977	masA	enolase-phosphatase E-1	42.3% ....
Aq1560	mglA1	gliding motility protein	42.4% ....
Aq1823	mglA2	gliding motility protein MglA	34.1% ....
Aq1789	mviB	'virulence factor' homolog MviB	29.7% ....
Aq587	neaC	N-ethylammeline chlorohydrolase	42.8% ....
Aq1820	nfeD	nodulation competitiveness protein NfeD	37.9% ....
Aq896	nifU	NifU protein	48.3% ....
Aq1300	omp	outer membrane protein	25.5% ....
Aq1507	omt	O-methyltransferase	39.5% ....
Aq967	ostA	organic solvent tolerance protein	22.0% ....
Aq141	pkcI	protein kinase C inhibitor (HIT family)	59.0% ....
Aq994	pncA	pyrazinamidase/nicotinamidase	39.1% ....
Aq057	sfsA	sugar fermentation stimulation protein	27.3% ....
Aq287	smb	small protein B	52.0% ....
Aq832	surE	stationary phase survival protein SurE	44.1% ....
Aq871	thdF	thiophene and furan oxidation protein	45.4% ....
Aq2021	tldD	TldD protein	40.9% ....
Aq773	tly	hemolysin	43.8% ....
Aq629	xcpC	chromosome assembly protein homolog	33.3% ....

ulate the activity of the histidine kinase CheA[28]. Phosphoryl groups from CheA are transferred to CheY, which then binds to the flagellar switch, altering the direction of flagellar rotation. Homologous chemotaxis systems are present in the archaea *Halobacterium salinarum*[29] and *Pyrococcus* sp. OT3 (H. Sizuya, personal communication), although the bacterial and archaeal flagellar apparatuses are not homologous[30]. The *M. jannaschii* genome also lacks homologues of known genes required for chemotaxis. Thus, either motility in *A. aeolicus* and *M. jannaschii* is undirected or input for controlling taxis is mediated through another, unidentified system. The most studied chemotaxis systems respond to sugars and amino acids, although responses to other inputs (for example, metals, redox potential, and light) may also occur. In contrast to all the organisms known to possess the classical chemotactic signal-transduction pathways, both *A. aeolicus* and *M. jannaschii* are obligate chemoautotrophs. Chemoautotrophs may respond to a different set of factors, such as concentrations of dissolved gas (CO_2, H_2 or O_2) or another critical parameter such as temperature.

In *E. coli*, the flagellar switch is essential for flagellar structure and function and coupling of chemotaxis signals. But the *A. aeolicus* genome encodes homologues of only two of the three *E. coli* proteins that make up the switch, FliG and FliN. Biochemical[31] and genetic[32] studies implicate the missing FliM protein as the receptor for phosphorylated CheY, the switch signal. The absence of both FliM and CheY in *A. aeolicus* supports the identification of FliM as the receptor for phosphorylated CheY in *E. coli*. This result also argues against a direct role for FliM in torque generation.

DNA replication and repair

The *A. aeolicus* primary replicative DNA polymerase, corresponding to the DNA polymerase III holoenzyme in *E. coli*, probably consists

Box 1 *Aquifex aeolicus* genome features

General

Length 1,551,335 bp
G + C content 43.4%
Protein-coding regions 93%
Stable RNA 0.8%
Non-coding repeats (none significant)
Intergenic sequences 6.2%

RNA

Ribosomal RNA	Chromosome coordinates
16S-23S-5S	572785-567770
16S-23S-5S	1192069-1197084

Transfer RNA
44 species (7 clusters, 28 single genes)

Other RNAs	Chromosome coordinates
tmRNA	1153844-1153498

Chromosomal coding sequences

849 similar to protein of known function (average length 1,066 bp)
256 similar to protein of unknown function (average length 896 bp)
407 unknown coding regions (average length 762 bp)
1,512 total (average length 956)

Extrachromosomal element (ECE)

Length 39,456 bp
G + C content 36.4%
Protein-coding regions 53.5%

ECE-coding sequences

1 similar to proteins of known function (length 948 bp)
4 similar to proteins of unknown function (average length 667 bp)
27 unknown coding regions (average length 648 bp)

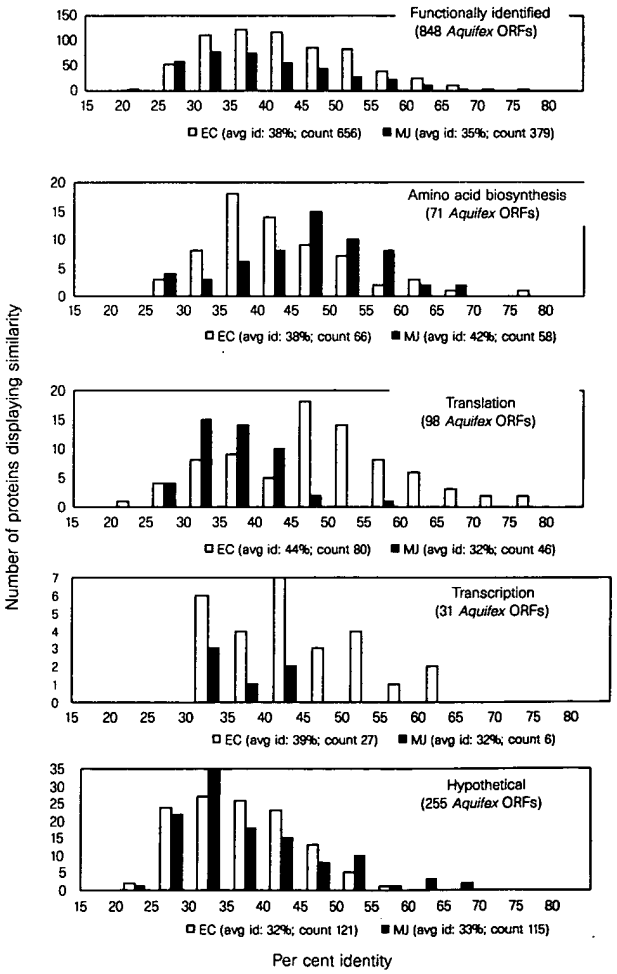


Figure 2 Histogram representation of the similarity of selected classes of predicted proteins to predicted proteins from the *E. coli* (EC) and *M. jannaschii* (MJ) genomes. Predicted *A. aeolicus* proteins representing each category were independently compared to sets of all potential polypeptides (≥100 amino acids) from the two genomes using FASTA[44]. If the top scoring alignment covered ≥80% of the length of the *A. aeolicus* protein, the score was plotted. There were more positives found in the *E. coli* genome in nearly every category. Hypothetical proteins (those identified by database match but of unknown function) are very similarly represented by *M. jannaschii* and *E. coli*. There are a small number of very highly conserved hypotheticals that are shared between *A. aeolicus* and *M. jannaschii*. Generally, biosynthetic categories show less discrimination than information-processing categories, which are clearly more *E. coli*-like. The variation in the apparent rates of evolution in different categories suggests that different phylogenies may be inferred depending on the sequence analysed. Within each graph, correspondence to *E. coli* is shown in white and *M. jannaschii* is shown in black. Avg id, average identity; count, number of proteins analysed.

of a core structure containing α- and ε-subunits, a γ-τ-subunit and an additional member of the γ-τ/δ′-family. A gene encoding a protein homologous to the β-sliding clamp was also found. This minimalistic complex lacks homologous θ-, δ-, χ- and ψ-subunits, as does the *Mycoplasma genitalium* holoenzyme[3]. Translation of the 54K (relative molecular mass) γ-τ-ATPase subunit may proceed without a programmed frameshift to produce a protein similar to the N-terminal region of the *E. coli* γ-subunit. DNA polymerase I is present as separate Klenow fragment and 5′ → 3′ exonuclease subunits, encoded by two non-adjacent ORFs. Although the repair polymerase, DNA polymerase II, has not been found in *A. aeolicus*, one ORF (Aq1422) encodes a protein similar to the eukaryotic DNA repair polymerase-β. A member of the same family has been identified in *Thermus aquaticus*[33] and *Bacillus subtilis*.

Transcriptional and translational apparatuses

The transcriptional apparatus of *A. aeolicus* is similar to that of *E. coli* and lacks any components specific to the Eukarya or Archaea (Fig. 2). In addition to the core RNA polymerase α-, β-, and β′-subunits, four σ-factors which determine promoter specificity are present (Table 1). Several different families of bacterial transcriptional regulators were also identified, including two-component systems. All of the ribosomal proteins and elongation factors common to other bacteria are present, indicating that all bacteria-specific ribosomal proteins were present in the common ancestor of *Aquifex* and other bacteria. Also present are the four *sel* genes required for the cotranslational incorporation of selenocysteine. These latter genes are clustered in a 15-kilobase-pair segment that also encodes the biosynthetic and structural proteins for formate dehydrogenase, the only selenocysteine-containing protein identified. The gene that encodes selenocysteine transfer RNA, *selC*, is apparently cotranscribed with the genes encoding the formate dehydrogenase structural proteins.

A. aeolicus lacks glutaminyl-tRNA and asparaginyl-tRNA synthetases. The genes required for transamidation of glutamyl-tRNA^{Gln} are present[34]. Charging of asparaginyl-tRNA is likely to proceed through the analogous reaction, as shown in halobacteria[35], although the genes(s) for that transamidase are unknown. The canonical methionyl- and leucyl-tRNA synthetases have only been seen previously as single polypeptide enzymes; however, in *A. aeolicus* the homologues appear fragmented into two subunits. In both cases, the genes that encode the N- and C-terminal portions are widely separated on the chromosome. No complete three-dimensional structural data are available for either methionyl- or leucyl-aminoacyl tRNA synthetases, but the subunit organization in the *A. aeolicus* aminoacyl-tRNA synthetases may reflect domain organization in the homologous proteins.

Thermophily

The *A. aeolicus* genome is the second completely sequenced genome of a hyperthermophile. By comparing the *A. aeolicus* and *M. jannaschii* genomes and contrasting them with the complete genomes of mesophiles, we can discover whether there are aspects of the genome or the encoded information that are diagnostic of hyperthermophiles. The G + C content of the stable RNAs is clearly indicative of the high growth temperature of the organism. This property can be used to identify stable RNAs against the relatively low G + C background of the *A. aeolicus* genome. The gene encoding tmRNA (or 10Sa RNA)[36], an RNA involved in tagging polypeptides translated from incomplete messenger RNAs for degradation, was located in this way.

Two genes for reverse gyrase are present in the genome. This is the only protein known to be present only in thermophiles. Other proteins, currently described as hypotheticals, may be diagnostic of hyperthermophiles but the data sets are not yet large enough to decide this with confidence.

Although features of stabilization may not be apparent in any given protein[37], a large enough data set may reveal general trends in amino-acid usage that are informative. Particularly important in this regard is inclusion of multiple genomes of hyperthermophiles so as not to allow the idiosyncracies of a single organism to bias the conclusions. As shown in Table 2, comparison of the amino-acid composition encoded by six genomes shows that use of individual amino acids can vary significantly from genome to genome. The data suggest trends that may be correlated with the thermostability of the encoded proteins. One apparent trend is that the hyperthemophile genomes encode higher levels of charged amino acids on average than mesophile genomes[38], primarily at the expense of uncharged polar residues. Glutamine in particular seems to be significantly discriminated against in the hyperthermophiles. Although this observation might be rationalized on the basis of

Table 2 Comparison of relative amino acid compositions (in percentages) of mesophiles and thermophiles

	Mesophiles				Thermophiles	
Amino acid	*H. influenzae*	*H. pylori*	*E. coli*	*Synechosystis*	*A. aeolicus*	*M. jannaschii*
A	8.21	6.83	9.55	9.07	5.90	5.54
C	1.03	1.09	1.11	1.01	0.79	1.27
D	4.98	4.77	5.20	5.07	4.32	5.52
E	6.48	6.88	5.91	6.20	9.63	8.67
F	4.46	5.41	3.87	3.75	5.13	4.20
G	6.65	5.76	7.42	7.77	6.75	6.41
H	2.05	2.12	2.26	1.93	1.54	1.43
I	7.10	7.20	5.95	6.31	7.32	10.45
K	6.32	8.94	4.48	4.26	9.40	10.36
L	10.50	11.18	10.56	10.93	10.57	9.38
M	2.44	2.28	2.86	2.12	1.92	2.33
N	4.89	5.83	3.88	3.76	3.60	5.24
P	3.72	3.28	4.41	5.09	4.07	3.38
Q	4.64	3.70	4.42	5.26	2.04	1.44
R	4.47	3.46	5.58	5.18	4.91	3.85
S	5.84	6.81	5.67	5.46	4.79	4.46
T	5.20	4.37	5.35	5.53	4.21	4.06
V	6.68	5.59	7.11	7.10	7.93	6.85
W	1.12	0.70	1.48	1.30	0.93	0.71
Y	3.12	3.68	2.83	2.78	4.13	4.33

	Mesophiles	Thermophiles
Charged residues (DEKRH)	24.11	29.84
Polar/uncharged residues (GSTNQYC)	31.15	26.79
Hydrophobic residues (LMIVWPAF)	44.74	43.36

an increased rate of deamidation of this residue at higher temperatures, asparagine does not appear subject to similar discrimination.

Phylogeny

The placement of the *Aquifex* lineage as one of the earliest divergences in the eubacterial tree[13,14] is interesting because of the insights it could provide into the ancestral eubacterial phenotype, including the hypothesized thermophilic nature of the first bacteria. Protein-based phylogenies often do not support the original rRNA-based placement[15,16,18]. Thus, the availability of some 1,500 genes from an *Aquifex* species would seem to offer a definitive resolution of the phylogeny. However, our analyses of ribosomal proteins, aminoacyl-tRNA synthetases, and other proteins do not do so, showing no consistent picture of the organism's phylogeny. We cannot make a more complete analysis and discussion here, but some observations can be made. These proteins do not yield a statistically significant placement of the *Aquifex* lineage or of other major eubacterial lineages. This situation partially reflects the inadequacy of some protein sequences as indicators of distant molecular genealogy because of their particular evolutionary dynamic, including the patterns and rates of amino-acid replacements. In some cases (such as the aminoacyl-tRNA synthetases for arginine, cysteine, histidine, proline and tyrosine), the analyses are further complicated by the presence of paralogous genes and/or apparent lateral gene transfers. It seems that a more extensive survey of genes and a better sampling of major eubacterial taxa will be required to confidently confirm or refute an early divergence of the *Aquifex* lineage.

Conclusions

Advances in sequencing techniques have allowed us to move beyond studies of single genes to studies of complete genomes only recently[2]. This rapid advance has created the opportunity to begin to characterize an organism with the full knowledge of the genome in hand. The complete genome summarized in this report represents our first view of *A. aeolicus*. The challenge now is to ask specific questions in ways which take advantage of the whole-genome data.

Beyond studies of any single organism in isolation, complete genomes allow comprehensive comparisons between organisms. For instance, comparisons of the similarity of genes can be made that reveal that genes in different categories vary in their relative conservation (Fig. 2). In addition, genome-wide trends are apparent. For example, why is there not more of a tendency to group functionally related genes (for example, biosynthetic pathways) into operons in *A. aeolicus*? This was also seen in the genome sequence of the autotroph *M. jannaschii*[1]. Is this because the autotrophic lifestyle decreases the need for selective regulation? There also seem to be a few multifunctional, fused proteins in *A. aeolicus* and *M. jannaschii*. Although this seems unlikely to be related to autotrophy, it might be associated with extreme thermophily. The large number of diverse genome sequences that will become available in the coming years will allow more detailed correlation of global genomic properties with particular physiologies. □

Methods

Sequencing strategy. The sequencing strategy used to assemble the complete genome was based on the whole genome random (or 'shotgun') approach, which has been successfully used for other genomes of similar size[1–4]. Shotgun sequencing projects are characterized by two phases: an initial completely random phase in which the bulk of the data is collected, followed by a closure phase where directed techniques are used to close gaps and complete the assembly. By pursuing a strategy where only 97% coverage was initially achieved, we were able to limit the number of sequences needed for the random phase to only 10,500 (ref. 39).

Sequences were generated from a small insert library constructed in λ ZAP II vectors[40,41] (average insert length 2.9 kilobase pairs). Two different methods were used for sequencing: first, dye-primer M13-21 and M13 reverse primer ABI Prism CS$^+$ ready reaction kits, analysed on 48-cm 4% polyacrylamide gels; and second, dye-terminator (ABI Prism FS+) reactions using two pBluescript-specific primers. These reactions were analysed on 36-cm 5% Long-Ranger gels.

The sequence fragments were assembled on an Apple Power Macintosh computer using Sequencher (Gene Codes, Ann Arbor, MI), an assembly and editing program. Assembly was typically performed in batches of roughly 200–400 sequences, and was followed by inspection and editing of the assemblies. All sequences in the set were compared with all others through this process. After assembly, the sequences comprised ~750 contigs at the end of the random phase. Sequences were obtained from both ends of ~200 randomly chosen clones from a fosmid library[42,43]. These sequences were then assembled with consensus sequences derived from the contigs of random-phase sequences using Sequencher. Gaps between contigs were closed by direct sequencing on fosmids not wholly contained within a contig. The fosmid library thus served a purpose analogous to that of the λ-scaffold in other projects[1–4]. The final eight gaps were closed by direct sequencing of polymerase chain reaction (PCR) products generated with the TaqPlus Long PCR System (Stratagene Cloning Systems, La Jolla, CA).

Consequences of reducing the number of sequences in the random phase are the large number of gaps that remain to be closed in the directed phase, and the reduction in overall coverage. To ensure that reduced coverage did not compromise accuracy, ~200 oligonucleotide primers were synthesized to resequence regions of ambiguity identified by visual inspection of the entire assembly. 13,785 sequences, with an average edited read length of 557 base pairs, constitute the final assembly. On the basis of a relatively small number of errors identified during the annotation process, we estimate the error frequency to be <0.01%, comparable to other published genomic sequence estimates.

Gene (ORF + RNA) identification and functional assignment approaches. Coding regions of the *A. aeolicus* genome were analysed and assigned using primarily the programs BLASTP[44] and FASTA[45] to search against a non-redundant protein database. Many analyses were carried out within the context of MAGPIE[46,47], an integrated computing environment for genome analysis. The results of these analyses are available for user interpretation, validation, and categorization. Additional ORFs were identified and start sites refined using the program CRITICA (J. H. Badger and G.J.O., unpublished program). Finally, all presumed 'intergenic regions' were examined with BLASTX for similarities to known protein sequences[48]. Transfer RNA genes were identified with the program tRNAscan-SE[49].

Received 26 August 1997; accepted 3 February 1998.

1. Bult, C. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).
2. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–511 (1995).
3. Fraser, C. M. *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403 (1995).
4. Tomb, J.-F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547 (1997).
5. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**, 4420–4449 (1996).
6. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC7803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136 (1996).
7. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
8. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546 (1996).
9. Huber, R. *et al.* *Aquifex pyrophilus* gen. nov. sp. nov. represents a novel group of marine hyperthermophilic hydrogen oxidizing bacteria. *Arch. Microbiol.* **15**, 340–351 (1992).
10. Reysenbach, L., Wickham, G. S. & Pace, N. R. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* **60**, 2113–2119 (1994).
11. Setchell, W. A. The upper temperature limits of life. *Science* **17**, 934–937 (1903).
12. Brock, T. D. The road to Yellowstone—and beyond. *Annu. Rev. Microbiol.* **49**, 1–28 (1995).
13. Burggraf, S., Olsen, G. J., Stetter, K. O. & Woese, C. R. A phylogenetic analysis of *Aquifex pyrophilus*. *Syst. Appl. Microbiol.* **15**, 353–356 (1992).
14. Pitulle, C. *et al.* Phylogenetic position of the genus *Hydrogenobacter*. *Int. J. Syst. Bacteriol.* **44**, 620–626 (1994).
15. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl Acad. Sci. USA* **93**, 7749–7754 (1996).
16. Klenk, H.-P., Palm, P. & Zillig, W. in *Molecular Biology of the Archaea* (eds Pfeifer, F., Palm, P. & Scleeifer, K. H.) 139–147 (Vch Pub, 1994).
17. Bocchetta, M. *et al.* Arrangement and nucleotide sequence of the gene (*fus*) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus*: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/*fus* sequences. *J. Mol. Evol.* **41**, 803–812 (1995).
18. Wetmur, J. G. *et al.* Cloning, sequencing, and expression of RecA proteins from three distantly related thermophilic eubacteria. *J. Biol. Chem.* **269**, 25928–25935 (1994).
19. Kawasumi, T., Igarashi, Y., Kodama, T. & Minoda, Y. *Hydrogenobacter thermophilus* gen. nov., sp. nov.

an extremely thermophilic, aerobic, hydrogen-oxidizing bacterium. *Int. J. Syst. Bacteriol.* **34,** 5–10 (1984).
20. Kristjannson, J., Ingason, A. & Alfredsson, G. A. Isolation of thermophilic obligately autotrophic hydrogen-oxidizing bacteria, similar to *Hydrogenobacter thermophilus*, from Icelandic hotsprings. *Arch. Microbiol.* **140,** 321–325 (1985).
21. Kryukov, V. R., Savel'eva, N. D. & Pusheva, M. A. *Calderobacterium hydrogenophilum* gen. nov., sp. nov. an extreme thermophilic bacterium and its hydrogenase activity. *Microbiology* (Engl. Trans. *Mikrobiologiya*) **52,** 611–618 (1983).
22. Riley, M. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57,** 862–952 (1993).
23. Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* **11,** 128–134 (1989).
24. Beh, M., Strauss, G., Huber, R., Stetter, K. O. & Fuchs, G. Enzymes of the reductive citric acid cycle in the autotrophic eubacterium *Aquifex pyrophilus* and in the archaebacterium *Thermoproteus neutrophilus*. *Arch. Microbiol.* **160,** 306–311 (1993).
25. Fuchs, G. in *Autotrophic Bacteria* (eds Schegel, H. G & Bowein, B.) 365–382 (Springer, New York, 1987).
26. Mai, X. & Adams, M. W. Characterization of a fourth type of 2-keto acid-oxidizing enzyme from a hyperthermophilic archaeon: 2-ketoglutarate ferredoxin oxidoreductase from *Thermococcus litoralis*. *J. Bacteriol.* **178,** 5890–5896 (1996).
27. Lim, J. H. *et al.* Cloning and expression of superoxide dismutase from *Aquifex pyrophilus*, a hyperthermophilic bacterium. *FEBS Lett.* **406,** 142–146 (1997).
28. Bourret, R. B., Borkovich, K. A. & Simon, M. I. Signal transduction pathways involving protein phosphorylation in prokaryotes. *Annu. Rev. Biochem.* **60,** 401–441 (1991).
29. Rudolph, J., Tolliday, N., Schmitt, C., Schuster, S. C. & Oesterhelt, D. Phosphorylation in halobacterial signal transduction. *EMBO J.* **14,** 4249–4257 (1995).
30. Jarrell, K. F., Bayley, D. P. & Kostyukova, A. S. The archaeal flagellum: a unique motility structure. *J. Bacteriol.* **178,** 5057–5064 (1996).
31. Welch, M., Oosawa, K., Aizawa, S. I. & Eisenbach, M. Effects of phosphorylation, Mg^{2+}, and conformation of the chemotaxis protein CheY on its binding to the flagellar switch protein FliM. *Biochemistry* **33,** 10470–10467 (1994).
32. Sockett, H., Yamaguchi, S., Kihara, M., Irikura, V. M. & Macnab, R. M. Molecular analysis of the flagellar switch protein FliM of *Salmonella typhimurium*. *J. Bacteriol.* **174,** 793–806 (1992).
33. Motoshima, H. *et al.* Molecular cloning and nucleotide sequence of the aminopeptidase T gene of *Thermus aquaticus* YT-1 and its high-level expression in *Escherichia coli*. *Agric. Biol. Chem.* **54,** 2385–2392 (1990).
34. Curnow, A. W. *et al.* Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl Acad. Sci. USA* **94,** 11819–11826 (1997).
35. Curnow, A. W., Ibba, M. & Söll, D. tRNA-dependent asparagine formation. *Nature* **382,** 589–590 (1996).
36. Tu, G. F., Reid, G. E., Zhang, J. G., Moritz, R. L. & Simpson, R. J. C-terminal extension of truncated proteins in *Escherichia coli* with a 10Sa decapeptide. *J. Biol. Chem.* **270,** 9322–9326 (1995).
37. Böhm, G. & Jaenicke, R. Relevance of sequence statistics for the properties of extremophilic proteins. *Int. J. Pept. Protein Res.* **43,** 97–106 (1994).
38. Choi, I.-G. *et al.* Random sequence analysis of genomic DNA of a hyperthermophile, *Aquifex pyrophilus*. *Extremophiles* **1,** 125–134 (1997).
39. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
40. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Res.* **16,** 7583–7600 (1988).
41. Alting-Mees, M. A. & Short, J. M. pBluescript II: gene mapping vectors. *Nucleic Acids Res.* **17,** 9494 (1989).
42. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89,** 8794–8797 (1992).
43. Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B. & Simon, M. I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* **20,** 1083–1085 (1992).
44. Altschul, S. F., Fish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
45. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448 (1988).
46. Gaasterland, T. & Sensen, C. W. MAGPIE: automated genome interpretation. *Trends Genet.* **12,** 76–78 (1996).
47. Gaasterland, T. & Sensen, C. W. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* **78,** 302–310 (1996).
48. Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nature Genet.* **3,** 266–272 (1993).
49. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

Acknowledgements. This work was supported in part by Department of Energy Microbial Genome Program grants (to R.V.S., C. R. Woese and G.J.O.). We thank C. Woese for his cooperation in the analysis of the genome and interest in the project; K. Stetter for continuing interest; G. Frey, J. Holaska, S. Peralta, D. Hafenbrandl, S. Delk, T. Robinson, and J. Arnett for technical assistance; and D. Robertson, J. Stein, I. Sanyal, T. Richardson, G. Hauska, and K. Williams for discussions.

Correspondence should be addressed to R.V.S. (e-mail: rswanson@diversa.com). Requests for *Aquifex aeolicus* should be addressed to R.H. (e-mail: Robert.huber@biologie.uni-regensburg.de). The sequences have been deposited with GenBank and assigned accession numbers AE000657 (chromosome) and AE000667 (extrachromosomal element).

STIC-ILL

From: Turner, Sharon
Sent: Thursday, June 17, 1999 4:35 PM
To: STIC-ILL
Subject: 09092297

Please RUSH!

Nature 1998, 392:353-58

Virology, 1992, 190:587-596

Mol Cell Biol, 1993, 13:1708-18

Sharon L. Turner, Ph.D.
CM1-8D08 GAU 1645
(703) 308-0056

VIROLOGY 190, 587–596 (1992)

Human Papillomavirus Type 13 and Pygmy Chimpanzee Papillomavirus Type 1: Comparison of the Genome Organizations[1]

MARC VAN RANST,*,† AKIRA FUSE,* PIERRE FITEN,* ERIK BEUKEN,* HERBERT PFISTER,‡ ROBERT D. BURK,† AND GHISLAIN OPDENAKKER*,[2]

*Laboratory of Molecular Medicine, Rega Institute, University of Leuven, Leuven B-3000, Belgium; †Department of Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, New York 10461; and ‡Institüt für Klinische Virologie, University of Erlangen-Nürnberg, Erlangen 8250, Germany

Received April 13, 1992; accepted June 12, 1992

Human papillomavirus type 13 (HPV-13) is associated with oral focal epithelial hyperplasia (FEH) in humans. A recent epidemic of a FEH-like disease in a pygmy chimpanzee (*Pan paniscus*) colony allowed us to clone a novel papillomavirus genome. To assess the homology between HPV-13 and the pygmy chimpanzee papillomavirus type 1 (PCPV-1), the complete nucleotide sequences of both FEH-related viruses were determined. In both viruses, all eight major open reading frames were located on one strand and the genomic organization was similar to that of other mucosal papillomaviruses. The genomes of PCPV-1 and HPV-13 showed extensive overall sequence homology (85%). They could be classified, using phylogenetic analysis, together with HPV types 6, 11, 43, and 44 in a group associated with benign orogenital lesions. These data indicate that two phylogenetically related papillomaviruses can elicit similar pathology in different primate host species, reflecting viral genomic similarities. © 1992 Academic Press, Inc.

INTRODUCTION

Papillomaviruses are a large and heterogeneous group of small icosahedral double-stranded DNA viruses that cause epithelial proliferations in a wide variety of higher vertebrates. To date, more than 65 different human papillomaviruses (HPVs) have been reported (de Villiers, 1989). At least 20 distinct HPV types are associated with anogenital premalignant and malignant lesions. Several benign epithelial tumors and hyperplasias in the oral cavity are also thought to be HPV related. These include verruca vulgaris, squamous cell papilloma, condyloma acuminatum, and focal epithelial hyperplasia (FEH, Heck's disease) (Syrjänen, 1987; Syrjänen *et al.*, 1987).

FEH is a well-defined clinical entity occurring only in the oral cavity. It was first described by Archard *et al.* (1965) in Navajo Indians in the United States. Clinically, FEH is characterized by multiple and discrete nodular elevations of the oral mucosa. The lesions can persist many years without extensive dissemination or malignant progression. Most FEH lesions tend to disappear spontaneously with few recurrences. The distribution of FEH is worldwide, but its prevalence is high (3.5 to 36%) in Indians in Central and South America and in Eskimos in Greenland and Alaska (Praetorius-Clausen, 1973). In contrast, the disorder is very rare in Caucasians, even if they live in areas where the disease is common among the native population. The epidemiology of FEH seems to indicate that it is a communicable disease. FEH occurs mainly in children and adolescents and often affects several members of the same family (Gomez *et al.*, 1969).

In 1983, Pfister *et al.* cloned and characterized human papillomavirus type 13 (HPV-13) DNA from a 13-year-old Turkish girl with FEH lesions. In 1987, another papillomavirus associated with FEH, HPV-32, was characterized by molecular cloning (Beaudenon *et al.*, 1987). Both viruses have also been identified recently in oral lesions of HIV-infected patients (Greenspan *et al.*, 1988).

In a recent epizootic outbreak of a FEH-like disease in a pygmy chimpanzee (*Pan paniscus*) colony in a zoological garden, we detected a papillomavirus that cross-hybridized to HPV-13 under stringent conditions and named it PCPV-1. This papillomavirus exhibited a divergent restriction endonuclease cleavage pattern, suggesting that it was similar but not identical to HPV-13 (Van Ranst *et al.*, 1991). Here we report the molecular cloning and characterization of PCPV-1 DNA. To study the relationship between these two papillomaviruses that cause similar diseases in different species, we also completely sequenced the genomes of both

[1] The nucleotide sequence data reported in this paper have been deposited with the GenBank/EMBL Data libraries under Accession Nos. X62843 and X62844.

[2] To whom requests for reprints should be addressed: Rega Institute for Medical Research, Minderbroedersstraat 10, B-3000 Leuven, Belgium.

0042-6822/92 $5.00

PCPV-1 and its human counterpart HPV-13. This enabled us to compare the genome structures and the phylogenetic relatedness with other papillomaviruses and to define all the viral proteins as well as specific regulatory sequences.

MATERIALS AND METHODS

Cloning of PCPV-1. For the cloning of PCPV-1, total cellular DNA was extracted from a monkey FEH lesion. Approximately 10 μg DNA was digested to completion with *Bam*HI restriction endonuclease, known to linearize the viral episomal DNA. The digested DNA was separated in a 1% low-melting-point agarose gel. The size-fractionated DNA of approximately 8000 bp was eluted from the gel and ligated in the *Bam*HI site of pUC119. Transformation of library efficiency *Escherichia coli* DH5α competent cells (GIBCO BRL Laboratories, Grand Island, NY) was done by the method of Hanahan (1983). After colony screening by direct DNA sequence analysis of recombinant plasmids, two plasmids were isolated and were shown to contain the PCPV-1 DNA in opposite orientations. All further procedures were performed following classical technology (Sambrook *et al.*, 1989).

DNA sequencing. The HPV-13 DNA cloned in pBR322 was digested with *Bam*HI and three fragments of 750, 1500, and 5000 bp were subcloned in the *Bam*HI site of pUC119 or pREGA (a pUC-derived plasmid vector suitable for both enzymatical and chemical sequencing). Further subclones were generated by cloning *Hin*dIII, *Eco*RI, *Pst*I, and *Xba*I fragments, and were used for sequencing with the universal M13 forward and reverse primers. On the basis of the partial sequence information, synthetic oligonucleotide primers were used for primer walking. Oligonucleotides were synthesized on an ABI 381A apparatus (Applied Biosystems, Inc., Foster City, CA) with the phosphoramidite protocol. Prior to use, the oligonucleotides were FPLC-purified on an anion-exchange Mono Q column (Pharmacia LKB Biotechnologies, Uppsala, Sweden), desalted, and lyophilized.

DNA sequencing was done on both strands on an Automated Laser Fluorescent (A.L.F.) DNA Sequencer (Pharmacia LKB Biotechnologies, Uppsala, Sweden). Single fluorescein labeled universal M13 primers or specific internal primers were annealed to the template DNA and enzymatically extended with T7 polymerase. The reactions were terminated with dideoxy-NTPs according to standard procedures provided by the manufacturer.

DNA sequence analysis and phylogenetic analysis. Assembly, analysis, and comparison of the nucleotide and amino acid sequences were computer assisted using the PC/GENE Software Package Release 6.6 (Intelligenetics, Inc., Mountain View, CA). Most HPV sequences used in the comparisons were downloaded from public computer databanks at the European Molecular Biology Laboratories (EMBL, Heidelberg, Germany) or at Genbank (Los Alamos National Laboratory, Los Alamos, NM). Pairwise alignments of nucleotide and amino acid sequences were performed using the method of Myers and Miller (1988), and multiple sequence alignments, using the CLUSTAL program (Higgins and Sharp, 1988). The computer alignments were corrected manually where appropriate. Phylogenetic analysis was done using parsimony algorithms in the PAUP software package (Release 3.0) (Swofford, 1990).

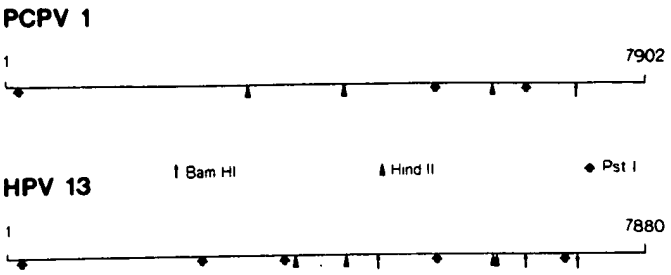


FIG. 1. Restriction maps for HPV-13 and PCPV-1 DNAs as inferred from the nucleotide sequences.

RESULTS

Restriction Analysis of HPV-13 and PCPV-1

Restriction enzyme digestion and Southern blot hybridization of DNA extracted from FEH biopsies revealed that HPV-13 had three *Bam*HI cleavage sites. HPV-13 had previously been cloned by partial digestion with *Bam*HI, and religation into bacteriophage λ (Pfister *et al.*, 1983). During sequencing, it became clear that the smallest of the three *Bam*HI partial digestion fragments (591 bp) had been religated in the opposite orientation. The restriction map for HPV-13 published by Pfister *et al.* (1983) proved to be correct, except for a fourth *Hin*dII site at position 6063 that would produce an extra 45-bp fragment, virtually undetectable by Southern blot analysis (Fig. 1). Based on the nucleotide sequence, a restriction endonuclease cleavage map was also prepared for PCPV-1 (Fig. 1). Although PCPV-1 and HPV-13 shared some restriction sites in homologuous positions, their restriction patterns were distinct.

General Features of the HPV-13 and PCPV-1 Nucleotide Sequences

The complete nucleotide sequences of HPV-13 and PCPV-1 were determined and are presented in Figs. 2

```
1    GTTTCTAACA ATCTTAAGTT TAAAAAATAG GTGGGACCGA AAACGGTTTT AACCGAAAAC GGTGATATAT AAACCAGCCC AAAAATTGAG CAAGCGGGGC
101  ATAATGGAAA GTGCAAATGC CTCCACGCCT GCAAAAACTA TAGACCAGTT GTGCAAGGAG TGCAACCTTT CTATGCACAG CTTGCAAATT CTATGCGTGT
201  TCTGCAGGAA AACCCTGTCC ACGGCAGAGG TTTATGCATT TCAGTATAAG AGTTTATATA TAGTGTGGCG AGGACAGTTT CCATTTGCGG CTTGTGCATG
301  CTGCTTAGAA ATACAAGGAA ACATTAACCA GTTTAGGCAT TTTGACTTCG CGGGATTTGC TGTAACAGTT GAAGAAGACA CAAAGCAGTC AATTTTGGAT
401  GTGCTAATTC GCTGCTATTT ATGCCACAAA CCATTGTGTG AAGTGGAGAA ACTAAGACAT ATTTTGCAGA AGGCACGATT TATTAAATTA AACAGCAGTT
501  GGAAAGGCCG CTGTTTTCAT TGCTGGTCAT CATGCATGGA AAATATCCTA CCTTAAAAGA CATTGTTTTA GAGCTGACTC CTGACCCTGT AGGTCTACAT
601  TGCAATGAGC AATTAGACAG CTCAGAAGAC GAGGTGGACG AACAAGCCAC GCAAGCCACG CAAGCCACGC AACATAGCAC ACTATTACAA TGCTACCAAA
701  TACTAACGTC CTGTAGTAAA TGTTGTAGCA ACGTCCGGCT GGTGGTGGAG TGTACAGGAC CTGACATTCA CGACCTACAC GACCTACTGC TGGGCACGCT
801  GAATATAGTG TGCCCTTTGT GTGCACCAAA AAGCTAACCA CGATGGCAGA GGATACAGGT ACTAATAATG AGGGGACGGG ATGCTCAGGA TGGTTTTTAG
901  TAGAGGCTGT AGTAGAACGA ACAACTGGGC AACAAATATC AGATGATGAG GATGAAACAG TGGAAGATAG TGGGTTGGAT ATGGTGGATT TCATAGATGA
1001 CAGACCTATT ACACACAATT CCGTGGAAGC ACAGGCATTG TTAAACGAGC AGGAGGCGGA TGCTCATTAT GCGGCTGTGC AGGACCTAAA ACGAAAGTAT
1101 TTAGGCAGTC CATATGTTAG TCCCCTAGGA CATGTTGAAC AGTCAGTGGA CTGTGATATA AGTCCTCGAT TGGACGCTAT AAAATTAAGT AGAAATTCTA
1201 AAAAAGTAAA GCGACGGCTG TTTCAATCAA GGGAAATAAC GGACAGTGGA TATGGCTATT CTGAAGTGGA AGCTGAAACG CAGGTAGAGA GAAATGGCGA
1301 ACCGGAAAAT GATTGTGGGG GTGGTGGACA CGGAAGGGAC AAAGAGGGGG AGGGACAGGT GCACACGGAA GTGCACACAG GCAGCCAGAT AGAAGAGCAC
1401 ACAGGGACCA CGCGGGTGTT AGAACTCCTT AAATGTAAGG ATGTAAGGGC TACATTGTAT GGTAAGTTTA AAGACTGTTA TGGGTTATCA TTTACAGATT
1501 TAATTAGACC ATTTAAAAGT GATAAAACAA CATGTGGGGA CTGGGTGGTT GCAGCATTTG GTATACATCA TAGTGTATCA GAGGCATTTG AAAAGTTAAT
1601 GCAGCCATTA ACAACATATA TGCATATACA ATGGCTTACA AATGCATGGG GGATGGTATT GTTAGTATTA ATAAGATTTA AAGTAAATAA AAGTAGATGC
1701 ACAGTGGCGC GAACACTGGC AACCTTTCTT AATATTCCTG AGGACCACAT GTTAATTGAA CCTCCCAAAA TACAAAGCAG TGTGGCAGCA TTATACTGGT
1801 TTAGAACAGG TATTTCTAAT GCTAGTATAG TAACTGGTGA AACACCAGAA TGGATAAAAA GGCAAACAAT TGTAGAGCAT GGACTTGCAG ATAATCAATT
1901 TAAATTAACT GAAATGGTGC AGTGGGCATA TGATAATGAT TTTTGTGATG AAAGCGAAAT AGCATTTGAA TATGCACAAC GAGGAGATTT TGATTCAAAT
2001 GCCAGGGCAT TTTTAAATAG TAATTGTCAG GCAAAATATG TAAAAGATTG TGCAACAATG TGCAAGCATT ATAAAAATGC AGAAATGAAA AAAATGTCTA
2101 TGAAACAATG GATAACATAT AGAAGTAAAA AAATAGAGGA AGCAGGAAAT TGGAAACCAA TAGTACAATT TTTAAGGCAT CAAAATATAG AATTTATTCC
2201 ATTTTTAAGT AAATTAAAAT TGTGGCTTCA TGGCACGCCA AAGAAAAACT GTATTGCAAT AGTGGGGCCA CCAGATACAG GCAAATCATG TTTTTGCATG
2301 AGCTTAATTA AGTTTTTAGG GGGCACAGTA ATTAGTTATG TAAATTCAAG TAGCCATTTT TGGCTGCAGC CATTATGTAA TGCAAAGGTA GCTTTGCTAG
2401 ATGATGCAAC GCAGTCATGC TGGGTATATA TGGACACATA CATGAGAAAT TTATTAGATG GCAATCCAAT GAGCATTGAT AGAAAACATA AGTCTTTAGC
2501 ATTAATAAAA TGTCCGCCAT TATTAGTAAC ATCTAATGTA GACATTACCA AAGATGACAA ATATAAATAT TTGTATAGTA GAGTAACAAC ACTTACATTT
2601 CCAAATCCAT TCCCTTTTGA CAGAAATGGG AATGCAGTAT ATGAGTTGTC TGATGCAAAC TGGAAATGTT TTTTTACAAG ATTGTCAGCA AGCCTAGATA
2701 TACAGGACTC TGAGGACGAG GACGATGGAG ACAATAGCCA AGCATTTAGA TGCGTGCCAG GAACAGTTGT TAGAACTGTA TGAAGAAAAT AGTAATGAAC
2801 TTAAAAAACA TATACAACAT TGGAAATGCT TAAGGTACGA AAGTGTACTC TTACACAAAG CACGCCAAAT GGGCCTAAGC CACATTGGAT TACAAGTGGT
2901 GCCACCATTG ACAGTATCAC AAGCTAAGGG ACATGAGGCA ATTGAAATGC AAATGACTTT AGAGACATTA CTAGAGTCTG AGTTTGGTAT GGAACCATGG
3001 ACTTTACAAG ATACAAGTCG TGAAATGTGG CTAACACCCC CAAAACGCTG TTTTAAGAAA CAGGGACAAA CTGTGGAAGT AAAATATGAC TGTAATACAG
3101 ACAATAGAAT GGATTATGTG TCGTGGACAT ACATATATGT GTTTGACACA GATAAATGGA CAAAGGTGAA AGGAATGGTA GATTATAAAG GGTTGTACTA
3201 CATACATGGA AATTTGAAAA CATATTATTT AGAGTTTGAA AAGGAGGCTA AAAAATATGG GGAAACGTTA CAATGGGAAG TATGTATTGG CAGCACAGTC
3301 ATATGTTCTC CTGCATCTGT ATCTAGTACT GTACAAGAAG TATCCATTGC TGGGCCTGCT TCATACTCCA CCACCACCTC CACACAGGCC TCCACCGCAG
3401 TGTCCTGCAG CGCCTCGGAA GAATGTGTGC AAGCGCCGCC TTGTAAACGA CAACGAGGAC GTTCACGTCC CATTGGAAAC CCCCAGAACA CACAAAGCAT
3501 TGTGTGTGTC ACAGACTACG ACACCCTGGA CAGTGCAAAC AACAACATCA ACGTTAACCA TTACAACAAT AACAAAGGAC GGGACAACAG TTACTGTGCA
3601 GCTACACCTA TAGTTCAATT ACAAGGTGAC TCTAATTGTC TAAAGTGTTT TCGATATAGA TTACATGAAA AATATAAAGA TTTATTTTTG TTAGCATCAT
3701 CTACATGGCA TTGGACCGCC CCTAATAATT CACAAAAACA TGCACTGGTA ACCTTAACCT ATGTAAATGA ACAACAAAGA CAAGACTTTT TAAAAACTGT
3801 AAAAATACCT CCAACCATAA CACATAAACT AGGTTTTATG TCATTGCAAT TGTTATAACA GCATATATTG TATGTAAATA TTTGTTGTGT GTGTGTATAT
3901 ATTGTAAATG GAATTTATAC CTGTGGATGT TAGTACACAG GCAACCAGCA AGTCATTACT GCCACTTGTA ATTGCACTTA CAGTGTGTGT AGTTAGCATT
4001 ATAACAATAT TGTGCATATC AGAGTTCTTG GTGTACACAA ACGTTTTAGT ACTAACATTA ATTTTATATG TACTTTTGTG GCTTTTACTA ACAACTCCCT
4101 TGCAATTCTA TTTACTAACC CTGTCTCTTT GCTTTCTTCC TGCGTTGTGT GTACACCAAT ATATTTTACA AACACAAGAA TAACTATACA CAATGTTAAC
4201 CTGTACTTTT GATGATGGTG ACACATGGTT GCTATTATGG TTAATTTTAT CATTTATTGT AGCCATTCTA GGGTTACTGT TGCTGTATAT AAGAACTGGA
4301 CATATGCATT GCCAGTGCTG GAGTAAATAA GTGGTTTTAT ATTTTGTGTG TATTCATTTA ATTATGGCAC ATAGTAGGGC TCGCAGACGC AAACGCGCTT
4401 CAGCTACACA ACTATATCAA ACTTGTAAGG CTTCTGGAAC ATGTCCTCCT GATGTTATAC CAAAGGTTGA ACAAAACACT CTTGCAGATA AAATATTAAA
4501 GTGGGGCAGT TTAGGAGTAT TTTTTGGGGG GCTTGGCATT GGCACAGGCT CTGGTACTGG CGGTAGGACT GGCTATGTAC CAGTAGGATC CACCCCACGC
4601 CCTGCCATAT CAACTGGGCC TACTGCACGT CCTCCTATTG TTGTTGATAC TGTTGGGCCT ACAGACCCTT CTATTGTATC TTTGGTAGAG GAATCAGCTA
4701 TTATTAATTC TGGAGTACCT GACCCTTTGC CTCCCGTTCA TGGGGGTTTT GAAATCACCA CATCTCAATC AGCCACTCCA GCAATATTGG ATGTGTCTGT
4801 TACAACACAA AACACTACGT CCACAAGTAT ATTTAGAAAT CCTGTTTTTT CAGAACCTTC TATTACACAA TCTCAACCTT CTATTGAAAG TGGTGCACAC
4901 GTGTTTATAT CGCCATCTAC TATTTCCCCT CATTCTACAG AAGACATTCC TTTAGATACA TTTATTGTAT CTTCCTCAGA TAGTAATCCT GCATCAAGCA
5001 CCCCTGTTCC AGCAACTGTT GCACGTCCAC GTCTAGGCCT TTACAGTAGG GCCTTACATC AAGTACAGGT TACTGATCCT GCCTTTTTAT CGTCGCCCCA
5101 ACGCCTTATA ACCTTTGATA ACCCTACATA TGAAGGTGAA GATATAAGTT TGCAGTTTGC ACACAATACC ATTCATGAAC CCCCTGATGA GGCATTTATG
5201 GATATTATAA GACTACATAG GCCAGCCATA ACATCACGGC GTGGTCTTGT TAGGTTTAGT AGAATTGGTC AGAGGGGGTC TATGTATACT CGAAGCGGCA
5301 AGCATATAGG TGGAAGGGTC CATTTCTTTA AGGATATTTC TCCTATATCT GCAGCTGCAG AAGAAATAGA ATTACACCCC CTTGTGGCTG CTGCACAGGA
5401 TCACAGTGGT TTGTTTGATA TTTATGCAGA ACCTGACCCT GACCCTGTGG CTGTAAACAC CTCTGGGTCA TTGTCTTCTG CCTCCACACC ATTTGCACAA
5501 TCTTCTTTGT CTTCCGCCCC ATGGGGTAAT ACTACTGTTC CTCTTTCACT ACCAGGTGAT ATATTTATAC AGCCTGGTCC TGACATAACA TTCCCAACTG
5601 CACCTACAGT AACGCCTTAT AATCCTGTTA CGCCTGCTTT ACCTACAGGT CCTGTTTTTA TTACTGCTTC TGGATTTTAT TTATATCCTA CATGGTATTT
5701 TACACGCAAA CGCCGTAAAC GTGTTTCCTT GTTTTTTACA GATGTGGCGG CCTAGTGACA ACAAACTATA TGTGCCTCCT CCCGCCCCTG TATCAAAAGT
5801 AATTACTACG GATGCCTATG TTACACGTAC CAACATATTT TATCATGCTA GCAGTTCTAG ACTACTTGCA GTGGGAAATC CTTATTTTCC TATTAAGAAA
5901 CAAAACAAAA CTGTTGTCCC TAAGGTATCT GGTTATCAGT TTAGGGTATT TAAAGTTGTA TTACCTGACC CTAATAAATT TGCCCTGCCT GACACATCTA
6001 TATTTGACTC AACTAGTCAA CGCTTAGTGT GGGCCTGTAC AGGTTTAGAG GTTGGTAGGG GTCAACCCTT AGGTGTTGGT ATTAGTGGTC ATCCATTATT
6101 AAATAAATAT GATGATGTGG AAAATTCTGC AAGTTATGCT GCCAATCCTG GTCAGGATAA TAGGGTTAAT GTGGCCATGG ACTATAAACA AACACAGTTA
6201 TGTTTAGTGG GCTGTGCACC TCCTTTAGGT GAACATTGGG GACAGGGCAA GCAATGTACT GGTGTAAATG TACAACCTGG AGATTGCCCT CCTTTAGAAT
6301 TAATTAGTAG TGTAATTCAG GATGGTGACA TGGTGGATAC AGGATTTGGA GCCATGAATT TTGCGGAATT GCAATCTAAT AAATCTGATG TGCCACTAGA
6401 CATATGCACG TCCACATGCA AATATCCTGA CTATTTACAA ATGGCTGCGG ATCCTTATGG AGACAGATTA TTTTTTTATC TGCGAAAGGA ACAAATGTTT
6501 GCAAGGCATT TCTTTAACAG GGCAGGCTCT GTTGGTGAAC AAATCCCAGC AGAATTATAT GTTAAGGGTA GTAATACACT TTCTAATAGT ATTTACTATA
6601 ATACTCCCAG TGGCTCTCTT GTGTCTTCTG AGGCCCAGTT GTTTAATAAA CCTTATTGGT TACAAAAGGC CCAGGGACAC AATAATGGTA TATGTTGGGG
6701 CAATCACTTG TTTGTTACTG TAGTTGATAC TACACGCAGT ACTAACATGA CTGTGTGTGC AGCCACTACA TCATCTCTTT CAGACACATA TAAGGCCACA
6801 GAATATAAAC AGTACATGCG ACATGTAGAA GAATTTGATT TACAATTTAT TTTTCAATTG TGCACTATTA AATTAACTGC AGAGGTTATG TCATATATTC
6901 ATACTATGAA TCCTACAATT CTAGAAGACT GGAACTTTGG GCTATCTCCC CCTCCTAATG GAACATTAGA AGACACATAT AGATATGTAC AATCTCAGGC
7001 CATAACGTGT CAAAAGCCTA CACCTGATAA AGAAAAACAG GATCCGTATG CGGGTCTTAG TTTTTGGGAG GTTAATCTTA AGGAAAAGTT TTCTAGTGAA
7101 CTAGATCAGT ATCCCCTTGG CAGAAAGTTT TTATTACAAA CAGGCGTTCA GTCTAGGTCC CCTATTCGTG TAGGTAGGAA ACGTGCTGCA TCTACATCTA
7201 CTGCCACACC TACTACACGT AAAAAAGCTA AAAGGAAATA ATAGTTTGTT TATGATTGTG TATGTATGTC ACGTTTGTTT GTACTGTATG TATGTTGTGT
7301 ACTGTATGTG TAATGTTGTA TGTATGTGCA TGTTACTTAT TAAAGAATGT GTGTGTGTGT TTGTATGCAA TAAATCTAAT CTGTGGTGTC CTGTTCCACC
7401 CTATGAGTAA GTGGTATGTT GTGTCTCGTG TGGTGTTTTG TATACTATAC TATAACATTA GTGCAACCAT TTTGTAACTT TTCTTACATT TTACGTCTCC
7501 ATATTAAGTG CAACCGATTT CGGTTGCTAT TGTTTCTGCG ACCGATTTGT TGCAGCACGC TGTTTATATA ATCTTACCTA CCGCCTGCCA AAATTATCCA
7601 CCGCTTGCCA AAATCACCCA CACACCTGGC GTTGCTAGGG CGCGGTTATA TATATTTACT AAATCTTACT AATCTTTCTA TCACTCATTT TACCTTTATA
7701 ACAATACTTT TGCTTTTCAA GTACATTTTT GTACTTACTA GCCAATGCCT GAAAGGTTTT TTGGCTACCA GCACTACATT TTTGTACAGT TAATGTTACA
7801 TGTATAAAAT GAGTAACCTA AGGTCACACA CCTGCAAACC GGTATCGGTT AAAACACACC CTCTATAGTT CCTTATAATT
```

FIG. 2. Complete nucleotide sequence of the coding strand of HPV-13 DNA (GenBank/EMBL Data Library Accession No. X62843). Position 1 was determined by alignment with the sequences of HPV-6 and HPV-11.

```
1     TTTAATAATA ATATCCTGTT TAAAAAATAG GAGGGACCGA AAACGGTTTC AACCGAAATC GGTGATATAT AAACCAGCCC ACAAATTAAG CAAGCGGGGC
101   ATAATGGAAA AAGCTAATGC CTCCACGTCC GCAAAAACGA TAGACCAGTT GTGCAAGGAG TGCAACCTTT GTATGCACAG TTTGCAAATT TTATGCGTGT
201   TTTGCAGGAA AACCCTGTCT ACTGCAGAGG TGTATGCATT TCAGTATAAG GATTTAAATA TTGTGTGGCA GGGCAATTTT CCATTTGCGG CCTGTGCATG
301   TTGTTTAGAA ATACAAGGAA AAGTTAATCA ATACAGGCAT TTTGACTTTG CTGCATATGC TGTAACAGTC GAAGAAGAAA TAAACAAGTC AATTTTTGAT
401   GTGAGAATTC GATGCTATTT GTGCCACAAA CCTTTGTGTG ACGTGGAAAA ATTACGACAC ATCTTGGAGA AAGCAAGATT CATTAAGTTA AACTGCGAGT
501   GGAAAGGGCG CTGCTTCCAT TGCTGGACAT CATGCATGGA AAATATACTA CCTTAAAGGA CATTGTGTTA GACCTAAGTC CTGACCCTGT AGGTCTACAT
601   TGCAATGAGC AATTAGACAG CTCAGAAGAA GATGAGGTGG ATGAACAAGC CACGCAAGCC ACGCAAGCCA CGTTCACACA ACATTACCAA ATAGTAACCT
701   GTTGTGGTCA GTGTGACAGC AACGTGCGTT TGGTTGTGGA CTGTACAGGA TCGGACATTC AACACCTACA CAAGCTGCTG CTGGGCTCAT TAAATATAGT
801   GTGCCCCTTG TGTGCTCCAC AAACCTAACC AGGATGGCCG ACAACACAGG TACAGACAAC AAGGGTACGG GGTGCTCAGG ATGGTTTTTA GTAGAGGCTA
901   TAGTAGACAG GAAAACTGGA GAAGAAATAT CAGATGATGA GGATGAAACA GTGGAAGATA GTGGGTTGGA TATGGTGGAT TTTATAGATG ACAGGTGTAT
1001  TACACACAAT TCCTTGGAAG CACAGGCATT GTTAAACGAG CAGGAGGCGG ATGCTCATTA TGCAGCTGTG CAAGACCTAA AACGAAAGTA TTTAGGTAGT
1101  CCCTATGTTA GTCCATTAGG CCATATTGAA CAGTCAGTCG AGTGTGATAT AAGTCCTCGA TTGAACGCCA TACAATTAAG TAGAAAACCT AAAAAAGTAA
1201  AGCGGCGGCT GTTCCAATCA AGGGAAATAA CGGACAGTGG ATATGGCCAT ACTGAAGTGG AAGTGGAAGC AGCAACGCAG GTAGAAAGAC ATGGCGAACC
1301  GGAAAATGGC TGTGGGGGGG GTGGACACGG AAGGGACAAA GAGGGGGAGG GACAGGTGCA TACGGAAGTG CACACAGAAA GCGAGATAGA ACACCACACG
1401  GGTACTACGC GGGTACTAGA ACTACTAAAA TGTAAGGATA TAAGGGCTAC ATTGCATGGT AAGTTTAAAC AATGCTATGG GCTATCGTTT ACAGATTTAA
1501  TTAGACAATT TAAAAGTAAT AAAACAACAT GTGAGGACTG GGTGGTGCGA GCATTTGGTG TGCATCATAG TGTGTCTGAG GCATTTGAAA AGTTAATACA
1601  GCCATTAACT ATATATAGGC ACATACAATG GCTAACAAAT GAATGGGGAA TGCTATTGTT AGTATTACTA AGATTTAAAG TAAATAAAAA TAGATGTACA
1701  GTAGCACGAA CACTGGCAAC ATTGCTTAAC ATTCCAGAAG ACCACATGCT AATTGAACCT CCAAAAATAC AAAGCAGTGT GGCAGCATTA TACTGGTTTA
1801  GAACCAGTCT ATCTAATGCC AGTATAGTAA CAGGAGAAAC ACCTGAATGG ATAGCAAGGC AAACAATAGT AGAACATGGC CTTGCAGATA GTCAATTTAA
1901  ATTAACTGAA ATGGTGCAAT GGGCCTATGA CAATGATTAT TGTGATGAAT GCGACATAGC ATTTGAATAT GCAAAACGAG CTGATTTTGA TTCTAATGCC
2001  AAAGCTTTTT TAAATAGTAA TTGTCAAGCA AAATATGTAA AGGATTGTGC AACAATGTGT AAGCATTATA AAAATGCAGA AATGAAAAAA ATGACAATGA
2101  ATCAATGGAT AAAACATAGA AGCAAAAAAA TAGATGAAAC AGGTAATTGG AAACCAATAG TGCAATTTTT AAGGCACCAA AATATAGAAT TTATTTCGTT
2201  TTTAAGTAAA TTAAAGTTGT GGCTTCAGGG CACACCAAAA AAAAACTGTA TTGCAATAGT GGGGCCACCA GATACAGGCA AATCAATGTT TTGCATGAGT
2301  TTAATAAAAT TTTTAGGGGG AACTGTAATT AGTTATGTAA ATTCAAGCAG CCATTTTTGG CTGCAACCGT TATGTAATAC TAAAGTAGCT TTGCTAGATG
2401  ATGCAACACA TTCATGCTGG GGATATATGG ATACATATAT GAGAAATTTA TTAGATGGTA ACCCTATGAG TATAGATAGA AAACATAAAT CGTTAGCATT
2501  AATAAAATGT CCGCCATTAT TAGTAACATC TAATATAGAT ATTACAACAG AGGAAAAATA TAAATATTTG TATAGTAGGG TAACAGTATT TAAATTTCCA
2601  AATCCATTCC CTTTTGACAG AAATGGGAAT GCAGTATATG AGTTGTGTGA TGCAAACTGG AAATGTTTTT TTGCAAGATT ATCAGCAAGT TTAGATATAC
2701  AGGACTCAGA GGACGAGGAC GATGGAGACA CTAGCCAAGC ATTTAGATGC GTGCCAGGAA CAGTTGTTAG AACTGTATGA AGAAAATAGT AATGAACTTA
2801  CAAAACATAT ACAACATTGG AAATGTGTAC GGCACGAAAA TGTACTGTTA TACAAAGCAC GCCAAATGGG CCTAAGCCAT ATTGGACTAC AAGTTGTGCC
2901  ACCATTAAAA GTATCACAAA CTAAGGGACA TGAGGCAATT GAAATGCAAA TGACGTTAGA AACAGTGCTA AAGTCAGAGT ATGGTACGGA ACCATGGACG
3001  TTACAAGAGA CAAGTTTTGA AATGTGGTTA ACACCACCAA AACATTGTTT TAAAAAACAG GGACAAACTG TGGAAGTAAG ATATGACTGC AATGCAGAAA
3101  ATTCAATGCA TTATGTATTG TGGAAATACA TTTATGTGTG TGAAAATGAC AGATGGCAAA AGGTAAAAGG AATGGTAGAC ATTAAAGGAT TATACTATAT
3201  GGTTGGACAG TGTAAAACAT ATTATATAGA CTTTGAAAAG GAGGCTAAAC AATATAGTAA AACATTACAA TGGGAAGTAT GTTATGACAG CAAAGTTATA
3301  TGTTCTCCTG CATCTGTATC TAGTACTGTG CAAGAAGTAT CCATTGCTGG GCCTACTTCA CACTCCACAA CCACCCTTGC ACAGGCCACC TGCGCAGTGT
3401  CATCCATCGC CACAGAAGAT AGTGTGCAAG CGCCGCCCTA TAAACGACTT CGAGGACCCT CCCACTGTGC TCGAAAACTC CAAAACACCT CTAACATTGT
3501  GTGTGCCACG GACCGTGGAA CCTTGGACAG TGAAAACAAC ATCAACAATA ACAATTACAA CAACAACAAC CAGCAACGGA ACAACAGTAA CAGTAGTGGT
3601  ACACCTATAG TGCAATTACA AGGTGACTCA AATAACTTAA AGTGTTTCAG ATATAGATTG CATGACAAAT ATAAACATTT ATTTATGCTA GCATCGTCTA
3701  CATGGCATTG GACCGCCTCT AGCAATTCAA CAAAAAATGC AATTGTAACA TTAACATATG TGAATGAACA ACAAAGACAA GATTTTTTAA ATACTGTAAA
3801  AATACCTGCC ACTATAAAAC ATACATTAGG GTTTATGTCG TTTCAATTAT TGTAACACCA TGTATTGTAC ATAACATGTA TTGTGTATAA ATGCTGTAAA
3901  TGGAATTACA GGTTGTACCT GTAGATTTTA CTGCAAAAGC AACCAGTCAA TCATTGTTGC CACTATTAAT TGCTCTTACT GTATGTTTTC TCAGTATTAT
4001  AATACTTATA TTTGTATCTG AGTTTCTACT ATATTCATCT GTTCTAGTAC TAACCTTACT TTTATATCTG TTGTTGTGGC TTTTACTAAC TCCCCCCTTG
4101  CAGTTTTTTT TACTAACCCT GTCTTTGTGC TTTTTGCCTG CGTTTTGTAT ACATCAATAT ATTTTGCAAA CTCAGCAACA ATAACTATAC ACAATGTTAA
4201  CGTGTACATT TAATGATGGT GATACATGGT TGTTATTATG GCTTTTTGCA TCAGTGTTAG TTGCAATTGT TGGGTTACTT TTACTGCATA TTAGAACTGG
4301  AAATTGTAAT TCACGTTTTT GTTATAGTAA ATAAGTTTTT TTATAGGTTG TGTGTAGTGG TTTCCTTATG GCACATAGTA GGCCTCGCAG ACGCAAACGT
4401  GCGTCTGCTA CACAGTTATA TCAAACCTGC AAAGCTTCTG GTACATGTCC TCCTGATATT ATTGCAAAGG TAGAACAAAA TACTCTTGCA GATAAAATAT
4501  TAAAGTGGGG AAGTTTAGGA GTGTTTTTTG GGGGGCTTGG TATTGGCACA GGGTCTGGTA CTGGCGGTAG AACTGGATAT GTTCCTGTAC AGACTGCCCC
4601  ACGCCCTGCC ATACCCTTTG GGCCTACTGC ACGTCCTCCT ATTATTGTTG ATACAGTTGG CCCTAGTGAC TCGTCTATTG TATCCTTAGT AGAAGATTCA
4701  ACTATTATTA ATTCAGCAGC GTCTGACTTT GTGCCTCCTA TTCGTGAGGG ATTTGAAATA AGCACCTCTG AAACTACTAC TCCAGCCATT TTAGATGTAT
4801  CTGTAACAAC ACACAACACT ACCTCTACAA GTATATTTAA AAATCCTGCC TTTGCAGAAC CTTCTATTGT TCAGTCACAA CCCTCAGTTG AAGCAAGTGG
4901  ACACGTTCTT ACATCTACAT ATACATCTAC TATTTCCTCT CACTCTGTAG AAGACATTCC ATTGGATACT TTTATTGTAT CTTCATCAGA TAGTAATCCT
5001  GCATCCAGTA CTCCTGTGCC TACACCTGTG GCACGTCCAC GACTTGGCCT TTATAGTAAA GCGTTGCAGC AAGTGCAGGT TACTAATCCT GCCTTTTTAT
5101  CATCGCCACA ACGCCTTATA ACTTTTGATA ACCCTGCTTA TGAAGGTGAA GATATAAGTT TACAGTTTCA GCACAATACT ATACATAACC CTCCTGATGA
5201  CGCTTTTATG GATATTGTAA GATTACATAG GCCGGCTATA ACGTCTAGGC GTGGTATTGT TAGGTTTAGT AGAATTGGTC AGCGAGGGTC CATGTATACA
5301  CGCAGTGGCA AACATATTGG TGGACGGGTA CATTTTTATA CAGACATTTC TCCTATATCT GCAGCTGCAG AGGAACTTGA AATGCAGCCT CTTGTGGCTG
5401  CTGCGCAGGA TGACAGTGGA TTATTTGATG TTTATGTAGA CCCTACCCCT GGTCCTGTAG CAGTACAAAA TATGTCGTAT CCCTCCTCCA CATCGTTTGT
5501  ACGGTCCTCT ATGTTTACCA CTAAGTGGGG TAATACTACT GTCCCTTTAT CATTACCAAG TAACATATTT GCACAGCCAG GCCCTGATAT AATATTTCCC
5601  GCCGCCCCTG GTGTACCACC GTATAACCCT GTTATACCGT CTTTACCTAT AACTCCTATT TTTATTAGTG GGTCTCAATT TTATTTACAT CCTTCCTTAT
5701  ATCTTGCACG CAAACGTCGT AAACGTGTTT CCTTGTTTTT TGCAGATGTG GCGGCCTAGT GACAACAAAC TATATGTGCC TCCTCCCGCC CCTGTATCAA
5801  AAGTAATTAC TACGGATGCA TATGTTACAC GTACAAAAAT ATTTTATCAT GCTAGCAGTT CTAGACTACT TGCAGTTGGT AATCCCTATT TTCCTATTAG
5901  AAAGGGTAAC AAAACTATTG TTCCCAAGGT ATCTGGATTT CAGTTTAGAG TATTTAAAAT AGTATTACCT GACCCTAATA AATTTGCTTT GCCTGACACA
6001  TCTATATTTG ATTCTACTAG TCAACGTTTG GTGTGGGCCT GTATAGGTTT AGAGGTAGGT AGGGGTCAGC CATTAGGTGT TGGTATTAGT GGACATCCTT
6101  TGTTAAACAA ATTTGATGAT GTAGAAAATT CTGCTAGTTA TGCTGTAAAC CCTGGCCAGG ATAATAGGGT TAATGTAGCA ATGGACTATA AACAAACACA
6201  GTTATGTTTG GTAGGTTGTG CACCTCCTTT AGGGGAACAT TGGGGTAAAG GCACACAATG TTCAGGTGTA AGTGTACAAG ATGGCGATTG TCCTCCATTG
6301  GAATTAGTTA CTAGTGTAAT TCAGGATGGT GATATGGTTG ATACAGGCTT TGGAGCCATG GATTTTGCAC AATTACAATC TAATAAATCA GATGTACCTT
6401  TAGATATATG CACAGCTACT TGCAAATATC CTGACTATTT ACAAATGGCT GCAGATCCTT ATGGTGACAG ATTATTTTTT TCTTTACGTA AAGAACAAAT
6501  GTTTGCAAGA CATTTTTTTA ATAGGGCAGG GACTGTTGGT GAGCAAATAC CAGAAGACTT ATTGGTTAAA GGGACTACCT CACGAGCTAC TGTTTCTAGT
6601  ACTATTTATT TTAATACTCC CAGTGGCTCT CTTGTGTCTT CTGAAGCTCA GTTGTTTAAT AAACCTTATT GGTTACATAA GGCTCAGGGA CACAATAATG
6701  GCATATGTTG GGGCAATACT TTGTTTGTTA CTGTTGTAGA TACTACACGA AGTACAAACA TGACTGTTTG TGCTTCCACA ACATCGTCGC CTTCTGCCAC
6801  ATACACAGCT TCAGAATATA AACAATACAT GCGACATGTG GAAGAATTTG ATTTGCAATT TATTTTTCAA TTATGCAGTA TTAAATTAAC TGCTGAAGTA
6901  ATGGCATATA TTCATACTAT GAATCCTACA GTTTTAGAAG AATGGAACTT TGGATTATCT CCCCCTCCCA ATGGAACATT AGAAGACACA TATAGATATG
7001  TTCAGTCTCA GGCTATAACA TGTCAAAAAC CTACTCCTGA TAAGGAGAAA CAGGATCCTT ATGCAGGTCT CAGTTTTTGG GAAGTTAATC TTAAAGAAAA
7101  GTTTTCTAGT GAGCTAGATC AATATCCGCT TGGTAGAAAG TTTTTATTAC AAACAGCGCT TCAAACTACA TCCTTTGCTC GTGCAGGAAC AAAACGTGCT
7201  GCAAGTACAT CTTCATCTAC ACCTACTACA CGTAAACGGG TTAAACGGAA ATAGTTTGTT ATTTTGTGTG TATGTTTGTT ATGTATGTTT GTTGTTTATA
7301  TGTTGTATAG TATATGTGTA ATGTTATATG TATGTTGTTT TTATGTATTA AAGAATGTAT GTGTGTGTTT GTGTGTAATA AACTTAATCT GTGGTGTCCT
7401  GTTCCACCCT ATGACTAACT ATTTTGTTGT TTATGTTATG TTGTATATTA TATAGTATTT ATATAACTAT TTAAAAACAT ATATATTAAA TTATACTCCA
7501  TATTGAGTAC GACCGAATTC GGTTGTCTAC TTTTACATCC TGTAACCAAT TTGTTGCAGC ACGTTACTTA TATAATCTTA TCTACTGTCT GCCAAAAATA
7601  ATTTATCCTA CCACCTACAC ACCTGGCGTT ACCAGGGTGC GGTTTTTATA TTTACAAAAT GTTATTAATC TCTGTAACAG TCATTCCACT TTCTAACAAT
7701  CGTTTTGCTT TTCTAGTAAC TTTTTCATAC TTTGTTGCCA ATGCCTAAAA TCCTTTTTGC CTACCAGCAC TACATTTTTG TAAACTTACT GTTCATTTTA
7801  CAAAATGAGT AACCGTGTTT GTTTTACAAA CTGAGTAACC TACGGTCACA CACCTGCACA CCGGTATCGG TTAAAACACA CCCTGTCTAA TCCTTTATTA
7901  TA
```

FIG. 3. Complete nucleotide sequence of the coding strand of PCPV-1 DNA (GenBank/EMBL Data Library Accession No. X62844). Position 1 was determined by alignment with the sequences of HPV-6 and HPV-11.

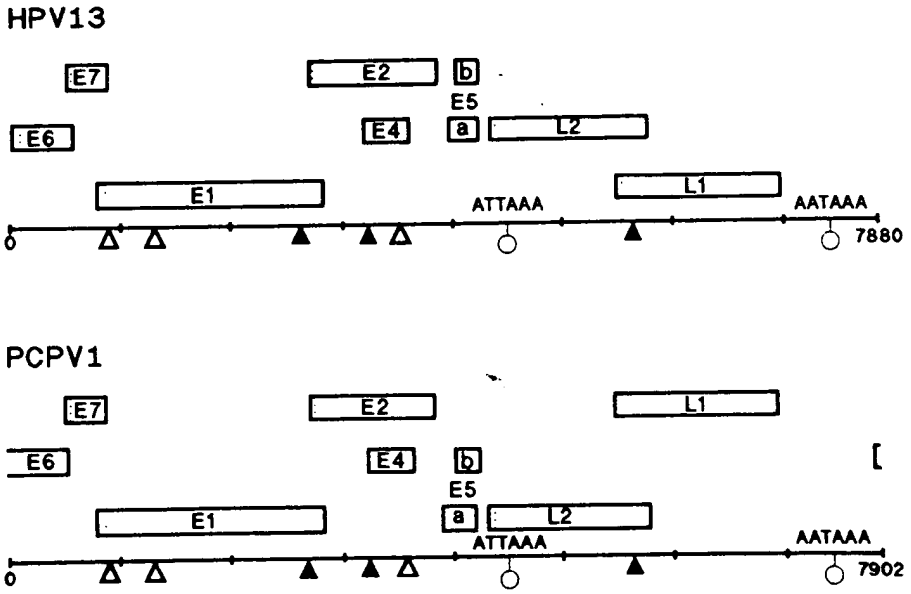


FIG. 4. Open reading frames (ORFs) of the genomes of HPV-13 and PCPV-1. The major open reading frames are located on one strand. The dotted vertical bar in the ORF represents the first ATG start codon. The scale indicates the distances along the linearized nucleotide sequences and the locations of the putative polyadenylation signals (ATTAAA or AATAAA, O). The splicing donor sites (Δ) and acceptor sites (▲) are indicated.

and 3, respectively. The DNA sequence of HPV-13 comprised 7880 bp and had a G + C content of 39.5%. PCPV-1 was 22 bp longer (7902 bp) and had a G + C content of 38%, the lowest G + C content of any sequenced nonhuman papillomavirus. The numbering for the HPV-13 and PCPV-1 nucleotide sequences was chosen to begin at a degenerate *Hpa*I site, homologous to nucleotide 1 of HPV types 6 and 11 (Schwarz *et al.*, 1983; Dartmann *et al.*, 1986). All eight major open reading frames (ORFs) of HPV-13 and PCPV-1 were located on one DNA strand. Similar to all sequenced papillomaviruses, except for HPV-39 (Volpers and Streeck, 1991), the complementary strand did not contain any significant ORFs and was therefore presumed to be noncoding. The genomic organization of the major ORFs in HPV-13 and PCPV-1 was very similar to that of the other mucosal papillomaviruses (Fig. 4). The exact location of the ORFs and their corresponding protein sequences are summarized in Table 1.

Overall Sequence Homology and Homology of the Different ORFs and Putative Proteins of HPV-13 and PCPV-1

Pairwise alignment of HPV-13 and PCPV-1 revealed 85% overall similarity, which was comparable to the 82% similarity between HPV-6 and HPV-11 (Dartmann *et al.*, 1986) and the 81% similarity between HPV-2 and HPV-57 (Hirsch-Behnam *et al.*, 1990) and between HPV-5 and HPV-47 (Zachow *et al.*, 1987; Kiyono *et al.*, 1990). Both HPV-13 and PCPV-1 closely resembled HPV types 6 and 11 (77–78%). The overall homology to HPV types 16 and 18 was around 60%, and that to cutaneous HPV types such as HPV-1, -5, -8, and -47 was lower than 50%.

The sequence similarity between HPV-13 and PCPV-1 was further demonstrated by pairwise comparison of the corresponding ORFs and their putative proteins (Table 2). When the ORFs of HPV-13 were compared to those of other mucosal papillomavirus types, highly significant homologies were found with HPV-6 and -11 in every ORF. We also examined the limited available DNA sequences of HPV-43 and HPV-44 and found that the HPV-13 and PCPV-1 E6 ORF showed the highest similarity to the HPV-44 E6 ORF. To further clarify the relations between the mucosal papillomaviruses, phylogenetic trees were constructed.

Phylogenetic Analysis

Phylogenetic trees can be constructed based on a comparative analysis of nucleotide or amino acid sequences and can be seen as a hypothetical representation of molecular evolutionary history. A phylogenetic tree can also be interpreted as an unbiased way of classifying multiple HPV types (Van Ranst *et al.*, manuscript in preparation). Four Cys–*X*–*X*–Cys repeats at invariant distances in the E6 gene of all papillomaviruses allow an unambiguous alignment, a prerequisite

TABLE 1

OPEN READING FRAMES OF HPV-13 AND PCPV-1 GENOMES

ORF	Virus	Nucleotide position			Number of bases	Number of amino acids[a]
		Start of ORF	First ATG	Stop codon		
E6	HPV-13	53	104	554	450	150
	PCPV-1	7892	104	554	450	150
E7	HPV-13	487	532	835	303	101
	PCPV-1	487	532	826	294	98
E1	HPV-13	765	843	2781	1938	646
	PCPV-1	717	834	2778	1944	648
E2	HPV-13	2698	2725	3856	1131	377
	PCPV-1	2695	2722	3853	1131	377
E4	HPV-13	3233	3257	3611	354	118
	PCPV-1	3230	3284	3608	324	108
E5	HPV-13	3908	3908	4181	273	91
	PCPV-1	3900	3900	4182	282	94
L2	HPV-13	4341	4364	5753	1389	463
	PCPV-1	4347	4368	5757	1389	463
L1	HPV-13	5613	5742	7239	1497	499
	PCPV-1	5653	5746	7252	1506	502

[a] Deduced from the first ATG from the start of the open reading frame (ORF).

for reliable phylogenetic evaluations. The phylogenetic tree of the E6 region of the mucosal papillomaviruses (Fig. 5) was rooted using HPV-1, a cutaneous HPV, as an outgroup. This analysis supported the notion that HPV-13 and PCPV-1 belong to the subgroup of the orogenital papillomaviruses together with HPV-6, -11, -43, and -44. These viruses are associated with condyloma acuminata and low-grade cervical neoplasia, rarely with cervical cancer. Since the predominantly genital HPV types 6 and 11 have also been found in oral condylomas and laryngeal papillomas, it is not surprising that HPV-13 has recently been detected outside of the oral cavity. HPV-13 was found in a case of low-grade cervical dysplasia using general primer-mediated polymerase chain reaction (PCR) (Snijders *et al.*, 1990), and in Bowenoid papulosis in an HIV-positive male by DNA *in situ* hybridization (Rolighed *et al.*, 1991). In both cases HPV-16 was also present in the lesion.

DISCUSSION

The present study describes the nucleotide sequences and genetic organizations of HPV-13, a type

TABLE 2

NUCLEOTIDE AND DEDUCED AMINO ACID SEQUENCE HOMOLOGIES OF OPEN READING FRAMES IN PCPV-1 AND SPECIFIC HUMAN PAPILLOMAVIRUS GENOMES

HPV-13 ORF	PCPV-1	HPV-6	HPV-11	HPV-16	HPV-18
E6	86 (86)[a]	74 (73)	75 (71)	59 (35)	52 (33)
E7	83 (76)	72 (67)	74 (71)	63 (52)	56 (42)
E1	90 (87)	81 (80)	81 (81)	64 (53)	65 (54)
E2	84 (76)	77 (65)	75 (67)	62 (45)	63 (39)
E4	78 (66)	78 (69)	75 (65)	60 (44)	59 (37)
E5	77 (ND)[b]	69 (ND)	69 (ND)	61 (ND)	60 (ND)
L2	81 (88)	71 (73)	72 (73)	62 (50)	63 (53)
L1	86 (88)	78 (82)	77 (81)	70 (69)	69 (63)
URR[c]	78 (NA)[d]	68 (NA)	65 (NA)	60 (NA)	64 (NA)

[a] Percentage nucleotide (amino acid) homology between the indicated HPV-13 ORF and the respective papillomavirus ORFs.
[b] ND, not determined.
[c] URR, upstream regulatory region.
[d] NA, Not applicable.

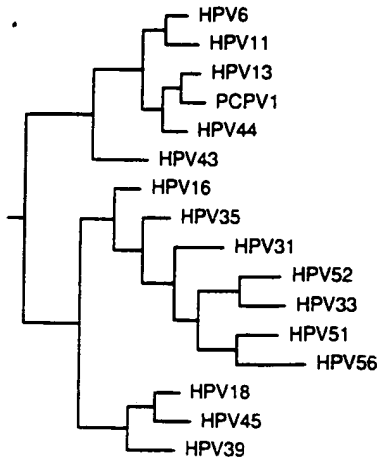


FIG. 5. Phylogenetic tree of 16 mucosal papillomaviruses. The branching order and branch lengths were inferred from the comparison of aligned E6 protein sequences using maximum parsimony algorithms in the PAUP program, Version 3.0 (Swofford, 1990). The tree was rooted using HPV-1, a cutaneous papillomavirus.

associated with a characteristic oral lesion in humans, and a pygmy chimpanzee virus that produces very similar oral lesions in that primate species. The two viruses were found to be closely related but distinguishable. This exemplifies how two related papillomaviruses produce similar pathological lesions in different species. The similarities between PCPV-1 and HPV-13 were investigated at the molecular genomic level.

Upstream Regulatory Region

The only major noncoding region (NCR) in papillomavirus genomes is the upstream regulatory region (URR). The URR was 748 bp long in HPV-13 and 757 bp long in PCPV-1. Approximately 125 bp downstream of the L1 stop codon (bp 7369 in HPV-13, bp 7377 in PCPV-1) this region contained the polyadenylation signal (AATAAA) for termination of mRNA transcripts from the late genes L1 and L2. This signal was followed at a distance of about 20 bp by a 20- to 40-bp region with a GT content of over 85%. A less frequently occurring polyadenylation signal (ATTAAA), positioned in the 3′ part of the L2 ORF (bp 4495 in HPV-13, bp 4499 in PCPV-1), might function as the termination signal for transcripts from the early genes. The same poly(A) signal configurations for early and late genes were found in corresponding positions in HPV-6 and HPV-11.

The URRs of HPV-13 and PCPV-1 contain four copies of the 12-bp palindromic E2 binding motif ($ACCN_6GGT$). In HPV-13 and PCPV-1, two E2 binding motifs were located between the "CAAT" and "TATA" boxes of the promoter proximal to the start codon of the E6 ORF. Binding of the E2 protein to such sites may sterically interfere with the binding of RNA polymerase and/or transcription factors, resulting in the repression of the transcription of the E6/E7 genes (Dostatni *et al.*, 1991; McBride *et al.*, 1991).

Five NFI sites were found in the URR of HPV-13 [nucleotides (nts) 7587, 7607, 7717, 7741, and 7761] and three sites in PCPV-1 (nts 7591, 7737, and 7757). Two putative sites for the transcription factor AP-1 (5′-Tt/gAGTCA-3′) were found (with a single mismatch) in the URR of HPV-13 and PCPV-1 at positions 7404, 7810 and 7812, 7838, respectively. In addition, a potential binding module for Oct-1 (5′-ATTTGCAT-3′) was also identified in HPV-13 (nt 7324).

Unlike the genital papillomaviruses, the URR of HPV-13 and PCPV-1 did not contain a glucocorticoid-responsive element (GRE) with sufficient homology to the consensus GRE sequence or GRE-like elements in other HPVs to be recognized as such. Therefore, HPV-13 and PCPV-1 may not be influenced by steroid hormones, in contrast to HPV-6 and -11. Papillomaviruses with a GRE are conceivably transcribed more efficiently in steroid-responsive cells, such as cervical cells. This might explain why HPV-13, albeit very similar to HPV-6 and -11, is only sporadically found in lesions of the genital mucosa.

E6

All papillomaviruses contain four Cys–*X*–*X*–Cys motifs spaced at regular and invariant intervals. The E6 gene product of both oncogenic and nononcogenic genital HPV types is involved in *in vitro* binding of the negative oncoprotein p53 (Scheffner *et al.*, 1990; Werness *et al.*, 1990). However, only the oncogenic types can subsequently promote the rapid degradation of p53 by the ubiquitin-dependent protease pathway (Crook *et al.*, 1991). The E6 proteins of HPV-13 and PCPV-1 are 150 amino acids long and their sequences are homologous to those of the E6 proteins of the nononcogenic group of papillomaviruses. HPV-13, HPV-44, and HPV-43 have one supplemental motif located 14 amino acids in front of the first conserved Cys–*X*–*X*–Cys. PCPV-1 contains two extra motifs 14 and 11 amino acids downstream of the first conserved motif.

In general, only the mucosal HPVs associated with malignant progression have splice donor/acceptor sites which may result in an internally spliced version of E6, referred to as E6* (Chow *et al.*, 1987; Schwarz *et al.*, 1987) and a E6–E7 colinear transcript. This splice site appears to be critical in the generation of a mRNA for E7 expression. In contrast, viruses lacking this splice donor/acceptor site transcribe the major colinear E6–E7 mRNA, responsible for generation of E7, from a promoter located within the E6 gene (Smotkin *et*

al., 1989). When comparing the different E6 sequences, it was found that HPV-13 and PCPV-1, together with HPV types 6, 11, and 44, do not have a splice donor/acceptor pair.

E7

The E7 protein contains a putative "cell-division" (cd) motif that is thought to mediate binding of the tumor suppressor protein pRB-105, gene product of the retinoblastoma gene (Goldsborough *et al.*, 1989; Dyson *et al.*, 1989). The rather complex pattern of this motif is (D,N)–L–*X*–C–*X*-(S,T,E)–X_{1-8}–(D,E)–(D,E,S,T)–(D/E). It is of interest to note that in PCPV-1 and HPV types 6, 11, and 13 the cd motif is degenerate with a glycine (G) as the first amino acid instead of an aspartic acid (D), as in the other papillomaviruses. PCPV-1 and HPV-13 contained two Cys–*X*–*X*–Cys motifs in the carboxy-terminal half of E7.

E1

The E1 gene is the largest and most conserved papillomavirus ORF and is involved in early replication events and episomal replication. E1 potentially interacts with an E2-encoded transcriptional activator protein, bound to the E2-responsive elements in the noncoding region (Mohr *et al.*, 1990). HPV-13 and PCPV-1 displayed up to 90% nucleic acid homology in the E1 ORF (Table 2). A stretch of 69 continuous identical nucleotides were found in the 3′ end of the E1 ORF, overlapping with the 5′ end of the E2 ORF.

E2

In HPV-13 and PCPV-1, a leucine zipper motif (L-X_6–L–X_6–L–X_6–L) (Landschulz *et al.*, 1988; Vinson *et al.*, 1989) was detected in the beginning of the carboxy-terminal part of E2. Furthermore, in HPV-13 and PCPV-1, an asparagine, needed to bend the putative alpha helixes and make their protruding sections fold around the target DNA, was found to be present at the correct position (18 amino acids upstream of the first leucine) in the middle of the DNA binding domain. Only HPV-13, PCPV-1, and HPV-11 contain a consensus leucine zipper domain in the E2 protein. HPV 6 has a phenylalanine instead of the first leucine. The high-risk papillomaviruses have more degenerate motifs.

It is tempting to speculate that a leucine zipper element in the *trans*-activator E2 protein could bring about the binding to the *cis*-E2-responsive elements in the URR. Binding of the E2 dimer to the motif proximal to the TATA box would sterically interfere with the positioning of the transcription complex at the promotor region in front of the E6 gene. Since not all amino acids in the leucine zipper region in papillomaviruses favor an alpha-helical conformation, it is necessary to clarify whether all these leucine zipper motifs have any functional significance, and whether the variations in these motifs can explain some of the biological differences between oncogenic and nononcogenic papillomaviruses.

E4 and E5

The E4 ORF overlaps with the E2 ORF. As in other papillomaviruses, the primary transcript encoding E4 is potentially spliced in HPV-13 and PCPV-1, generating a putative E4 mRNA preceded by the 5′ end of E1. In benign lesions, the E4 mRNA is the most abundant transcript of viral genes, and is thought to have a role in viral maturation.

The region between the end of the E2 ORF and the beginning of the L2 ORF contains a small noncoding region and an ill-conserved E5 ORF, coding for a hydrophopic protein. In HPV-13 and PCPV-1, and also in HPV-6 and HPV-11, this region is covered by two small ORFs, E5a and E5b.

L1 and L2

The late region of the papillomavirus genomes contains two large ORFs, L1 and L2, coding for the major and minor structural coat proteins. The L1 protein is evolutionarily highly conserved and has a large number of glycosylation sites, thought to stabilize the structure of the virus particle (Larsen *et al.*, 1987). The L1 ORFs of HPV-13 and PCPV-1 are very similar, and a stretch of 77 identical nucleotides was found in the 5′ end. The high degree of homology between the late proteins of HPV-13 and PCPV-1 is striking (Table 2). These proteins are likely to interact with species-specific and/or cell-type specific virus receptors. Such receptors might determine the viral host range and/or tissue specificity.

The close phylogenetic relationship between humans and pygmy chimpanzees and the extensive similarities between their respective FEH-related papillomaviruses raise the possibility of transmission across species barriers.

The reason why FEH is more common in specific populations or families (Praetorius-Clausen, 1973) is largely unknown, although one can presume that genetic factors play a role. This is not unlike epidermodysplasia verruciformis (EV), a rare autosomal recessive skin disease that predisposes patients to infection with EV-specific HPVs. FEH is very rare in Caucasians and Asiatics, but HPV-13 DNA has recently been demonstrated in oral and anogenital lesions in HIV-infected Caucasians (Greenspan *et al.*, 1988; Rolighed *et al.*, 1991). This could indicate that a reservoir of the virus

might be present in the general Caucasian population. Because of the lack of a sensitive assay, no data exist on the prevalence of the virus in subclinical infections in asymptomatic subjects. The possibility of detecting HPV-13 by PCR in a few cells, collected by scraping the oral mucosa with a wooden spatula, would enable epidemiological studies with little discomfort for the subjects.

Infections with papillomaviruses appear to be uncommon in nonhuman primates. This can be explained by the maintenance of monkeys in closed colonies in captivity, where pathogens are not easily transmitted between different communities. Alternatively, since papillomaviruses rarely cause life-threatening diseases, it may be due to a lack of careful examination. Papillomaviruses have been found in genital, cutaneous, and oral lesions in a number of nonhuman primates. In these species, the viruses are associated with similar histological manifestations as in humans. A venereal papillomavirus DNA was cloned from a lymph node metastasis of a primary penile squamous cell carcinoma in a rhesus monkey [*Maccaca mulatta;* rhesus papillomavirus type 1 (RhPV-1)] (Ostrow *et al.*, 1991a,b). A related papillomavirus was cloned from a papilloma on the penile shaft of an Abyssinian colobus monkey [*Colobus guereza kikuyuensis; Colobus guereza* papillomavirus type 1 (CgPV-1)] (O'Banion *et al.*, 1987; Reszka *et al.*, 1991).

Cutaneous papillomas were observed on the feet of a black and white colobus monkey (*Colobus polykomus*) and an Abyssinian colobus monkey (Boever and Kern, 1976; Rangan *et al.*, 1980). *Colobus guereza* papillomavirus type 2 (CgPV-2) was cloned from this lesion (Kloster *et al.*, 1988).

In this study, we report the characterization of an oral papillomavirus genome cloned from the pygmy chimpanzee, *Pan paniscus* (PCPV-1). Three case reports of FEH have been reported in the common chimpanzee, *Pan troglodytes,* a species evolutionarily highly related to the pygmy chimpanzee (Hollander and van Noord, 1972; Tate *et al.*, 1973; Glad and Nesland, 1986). Recently, a *P. troglodytes* papillomavirus (PtPV) was cloned from one of these FEH lesions (Favre *et al.*, personal communication).

As the complete DNA sequences of all these animal viruses become available for analysis, a more detailed picture of papillomavirus phylogeny and perhaps transmission will emerge.

ACKNOWLEDGMENTS

This research was supported by grants from the Cancer Fund of the General Savings and Retirement Fund (ASLK/CGER) and in part by NIH Grant 5P30 CA-13330. M.V.R. was supported by a fellowship from the Belgian American Educational Foundation and from the D. Collen Research Foundation. A.F. is a visiting professor at the University of Leuven. G.O. is associated with the Belgian NFWO and thanks the NFWO for the automated DNA sequencer through Grant 3.0027.90.

The authors thank Arthur Van Aerschot for help in oligonucleotide synthesis, J. Van Damme for critical comments, and Alfons Billiau for support.

REFERENCES

ARCHARD, H. O., HECK, J. W., and STANLEY, H. R. (1965). Focal epithelial hyperplasia: An unusual oral mucosal lesion found in Indian children. *Oral Surg. Oral Med. Oral Pathol.* **20,** 201–212.

BEAUDENON, S., PRAETORIUS, F., KREMSDORF, D., LUTZNER, M., WORSAAE, N., PEHAU-ARNAUDET, G., and ORTH, G. (1987). A new type of human papillomavirus associated with oral focal epithelial hyperplasia. *J. Invest. Dermatol.* **88,** 130–135.

BOEVER, W. J., and KERN, T. (1976). Papillomas in black and white colobus monkeys. *J. Wildl. Dis.* **12,** 180–181.

CHOW, L. T., NASSERI, M., WOLINSKY, S. M., and BROKER, T. R. (1987). Human papillomavirus type 6 and 11 mRNAs from genital condylomata acuminata. *J. Virol.* **61,** 2581–2588.

CROOK, T., TIDY, J. A., and VOUSDEN, K. H. (1991). Degradation of p53 can be targeted by HPV E6 sequences distinct from those required for p53 binding and trans-activation. *Cell* **67,** 547–556.

DARTMANN, K., SCHWARZ, E., GISSMANN, L., and ZUR HAUSEN, H. (1986). The nucleotide sequence and genome organization of human papilloma virus type 11. *Virology* **151,** 124–130.

DE VILLIERS, E.-M. (1989). Heterogeneity of the human papillomavirus group. *J. Virol.* **63,** 4898–4903.

DOSTATNI, N., LAMBERT, P. F., SOUSA, R., HAM, J., HOWLEY, P. M., and YANIV, M. (1991). The functional BPV-1 E2 trans-activating protein can act as a repressor by preventing formation of the initiation complex. *Genes Dev.* **5,** 1657–1671.

DYSON, N. P., HOWLEY, P. M., MUNGER, K., and HARLOW, E. (1989). The human papillomavirus 16 E7 oncoprotein is able to bind to the retinoblastoma gene product. *Science* **243,** 934–937.

GLAD, W. R., and NESLAND, J. M. (1986). Focal epithelial hyperplasia of the oral mucosa in two chimpanzees (*Pan troglodytes*). *Am. J. Prim.* **10,** 83–89.

GOLDSBOROUGH, M. D., DISILVESTRE, D., TEMPLE, G. F., and LORINCZ, A. T. (1989). Nucleotide sequence of human papillomavirus type 31: A cervical neoplasia-associated virus. *Virology* **171,** 306–311.

GOMEZ, A., CALLE, J., ARCILA, G., and PINDBORG, J. J. (1969). Focal epithelial hyperplasia in a half breed family of Colombians. *J. Am. Dent. Assoc.* **79,** 663–667.

GREENSPAN, D., DE VILLIERS, E.-M., GREENSPAN, J. S., DE SOUZA, Y. G., and ZUR HAUSEN, H. (1988). Unusual HPV types in oral warts in association with HIV infection. *J. Oral Pathol. Med.* **17,** 482–488.

HANAHAN, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166,** 557–580.

HIGGINS, D. G., and SHARP, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73,** 237–244.

HIRSCH-BEHNAM, A., DELIUS, H., and DE VILLIERS, E.-M. (1990). A comparative sequence analysis of two human papillomaviruses types 2a and 57. *Virus Res.* **18,** 81–98.

HOLLANDER, C. F., and VAN NOORD, M. J. (1972). Focal epithelial hyperplasia: A virus-induced oral mucosal lesion in the chimpanzee. *Oral Surg. Oral Med. Oral Pathol.* **33,** 220–226.

KIYONO, T., ADACHI, A., and ISHIBASHI, M. (1990). Genome organization and taxonomic position of human papillomavirus type 47 inferred from its DNA sequence. *Virology* **177,** 401–405.

KLOSTER, B. E., MANIAS, D. A., OSTROW, R. S., SHAVER, M. K.,

McPherson, S. W., Rangen, S. R. S., Uno, H., and Faras, A. J. (1988). Molecular cloning and characterization of the DNA of two papillomaviruses from monkeys. *Virology* **166,** 30–40.

Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. *Science* **240,** 1759–1764.

Larsen, P. M., Storgaard, L., and Fey, S. J. (1987). Proteins present in bovine papillomavirus particles. *J. Virol.* **61,** 3596–3601.

McBride, A. A., Romanczuk, H., and Howley, P. M. (1991). The papillomavirus E2 regulatory proteins. *J. Biol. Chem.* **266,** 18411–18414.

Mohr, I. J., Clark, R., Sun, S., Androphy, E. J., MacPherson, P., and Botchan, M. R. (1990). Targeting the E1 replication protein to the papillomavirus origin of replication by complex formation with the E2 transactivator. *Science* **250,** 1694–1699.

Myers, E. W., and Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4,** 11–17.

O'Banion, M. K., Sundberg, J. P., Shima, A. L., and Reichmann, M. E. (1987). Venereal papilloma and papillomavirus in a colobus monkey (*Colobus guereza*). *Intervirology* **28,** 232–237.

Ostrow, R. S., Labresh, K. V., and Faras, A. J. (1991a). Characterization of the complete RhPV-1 genomic sequence and an integration locus from a metastatic tumor. *Virology* **181,** 424–429.

Ostrow, R. S., McGlennen, R. C., Shaver, K., Kloster, B. E., Houser, D., and Faras, A. J. (1991b). A rhesus monkey model for sexual transmission of a papillomavirus isolated from a squamous cell carcinoma. *Proc. Natl. Acad. Sci. USA* **87,** 8170–8174.

Pfister, H., Hettich, I., Runne, U., Gissmann, L., and Chilf, G. N. (1983). Characterization of human papillomavirus type 13 from focal epithelial hyperplasia Heck lesions. *J. Virol.* **47,** 363–366.

Praetorius-Clausen, F. (1973). Geographical aspects of oral focal epithelial hyperplasia. *Pathol. Microbiol.* **39,** 204–213.

Rangan, S. R. S., Gutter, A., Baskin, G. B., and Anderson, D. (1980). Virus associated papillomas in colobus monkeys (*Colobus guereza*). *Lab. Anim. Sci.* **30,** 885–889.

Reszka, A. A., Sundberg, J. P., and Reichmann, M. E. (1991). *In vitro* transformation and molecular characterization of colobus monkey venereal papillomavirus DNA. *Virology* **181,** 787–792.

Roligheb, J., Sorensen, I. M., Jacobsen, N. O., and Lindeberg, H. (1991). The presence of HPV types 6/11, 13, 16 and 33 in bowenoid papulosis in an HIV-positive male, demonstrated by DNA *in situ* hybridization. *APMIS* **99,** 583–585.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J., and Howley, P. M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* **63,** 1129–1136.

Schwarz, E., Durst, M., Demankowski, C., Lattermann, O., Zech, R., Wolfsperger, E., Suhai, S., and Zur Hausen, H. (1983). DNA sequence and genome organization of genital human papillomavirus type 6b. *EMBO J.* **2,** 2341–2348.

Schwarz, E., Schneider-Gadicke, A., and Zur Hausen, H. (1987). Human papillomavirus type-18 transcription in cervical carcinoma cell lines and in human cell hybrids. *Cancer Cells* **5,** 47–53.

Smotkin, D., Prokoph, H., and Wettstein, F. O. (1989). Oncogenic and nononcogenic human genital papillomaviruses generate the E7 mRNA by different mechanisms. *J. Virol.* **63,** 1441–1447.

Snijders, P. J. F., van den Brule, A. J. C., Schrijnemakers, H. F. J., Snow, G., Meijer, C. J. L. M., and Walboomers, J. M. M. (1990). The use of general primers in the polymerase chain reaction permits the detection of a broad spectrum of human papillomavirus genotypes. *J. Gen. Virol.* **71,** 173–181.

Swofford, D. L. (1990). "PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0." Computer program distributed by the Illinois Natural History Survey, Champaign.

Syrjänen, S. M. (1987). Human papillomavirus infections in the oral cavity. *In* "Papillomaviruses and Human Disease" (K. Syrjänen, L. Gissmann, and L. G. Koss, Eds.), pp. 104–137. Springer Verlag, Berlin.

Syrjänen, S. M., Syrjänen, K. J., Happonen, R.-P., and Lamberg, M. A. (1987). *In situ* DNA hybridization analysis of human papillomavirus (HPV) sequences in benign oral mucosal lesions. *Arch. Dermatol. Res.* **279,** 543–549.

Tate, C. L., Conti, P. A., and Nero, E. P. (1973). Focal epithelial hyperplasia in the oral mucosa of a chimpanzee. *J. Am. Vet. Med.* **163,** 619–621.

Van Ranst, M., Fuse, A., Sobis, H., de Meurichy, W., Syrjänen, S. M., Billiau, A., and Opdenakker, G. (1991). A papillomavirus related to HPV type 13 in oral focal epithelial hyperplasia in the pygmy chimpanzee. *J. Oral Pathol. Med.* **20,** 325–331.

Vinson, C. R., Sigler, P. B., and McKnight, S. L. (1989). Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science* **246,** 911–916.

Volpers, C., and Streeck, R. E. (1991). Genome organization and nucleotide sequence of human papillomavirus type 39. *Virology* **181,** 419–423.

Werness, B. A., Levine, A. J., and Howley, P. M. (1990). Association of human papillomavirus types 16 and 18 E6 proteins with p53. *Science* **248,** 76–79.

Zachow, K. R., Ostrow, R. S., and Faras, A. J. (1987). Nucleotide sequence and genome organization of human papillomavirus type 5. *Virology* **158,** 251–254.